Structural analysis of metabolic networks

D I S S E R T A T I O N

zur Erlangung des akademischen Grades
doctor rerum naturalium
(dr. rer. nat.)
im Fach Biologie

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät I
Humboldt-Universität zu Berlin

von

Herr Dipl.-Phys. Oliver Ebenhöh

geboren am 5.2.1970 in Heidelberg

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr. Jürgen Mlynek

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät I:
Prof. Dr. Michael Linscheid

Gutachter:

1. Prof. Dr. Reinhart Heinrich
2. Priv.-Doz. Stefan Schuster
3. Prof. Dr. Werner Ebeling

eingereicht am:	13. Januar 2003
Tag der mündlichen Prüfung:	1. April 2003

Summary

This work is based on the hypothesis that present-day metabolic systems are the result of an evolutionary development which is governed by mutation mechanisms and natural selection principles. This concerns in particular the kinetic parameters of the enzymes as well as the stoichiometry of the system. Therefore, it can be assumed that these parameters have reached, during the course of their evolution, values which imply certain optimal properties with respect to their biological function.

Based on this hypothesis, the structural designs of metabolic reaction systems are investigated. Two models following this premise are presented and both models have as a common research objective the explanation of structural properties of metabolic systems using optimisation principles.

The first model which is covered in chapter 2 analyses the structural design of ATP and NADH producing systems such as glycolysis and the citric acid cycle (TCA). It is assumed that these pathways combined with oxidative phosphorylation have reached, during their evolution, a high efficiency with respect to ATP production rates. On the basis of kinetic and thermodynamic principles conclusions are derived concerning the optimal stoichiometry of such pathways. Extending previous investigations, both the concentrations of adenine nucleotides as well as nicotinamide adenine dinucleotides are considered variable quantities. This implies the consideration of the interaction of an ATP and NADH producing system, an ATP consuming system, a system coupling NADH consumption with ATP production, and a system consuming NADH decoupled from ATP production. It is examined in what respect real metabolic pathways can be considered optimal by studying a large number of alternative pathways. The kinetics of the individual reactions are described by linear or bilinear functions of reactant concentrations. In this manner, the steady state ATP production rate can be calculated for any possible ATP and NADH producing pathway. It is shown that most of the possible pathways result in a very low ATP production rate and that the very efficient pathways share common structural properties. Optimisation with respect to the ATP production rate is performed by an evolutionary algorithm. The following results of our analysis are in close correspondence to the real design of glycolysis and the TCA cycle: (i) In all efficient pathways the ATP consuming reactions are located near the beginning. (ii) In all efficient pathways NADH producing reactions as well as ATP producing reactions are located near the end. (iii) The number of NADH molecules produced by the consumption of one energy-rich molecule (glucose) amounts to four in all efficient pathways. A distance measure and a measure for the internal ordering of

reactions are introduced to study differences and similarities in the stoichiometries of metabolic pathways.

The second model described in chapter 3 follows a more general approach. Starting from a limited set of reactions describing changes in the carbon skeleton of biochemical compounds complete sets of metabolic networks are constructed. The networks are characterised by the number and types of participating reactions. Elementary networks are defined by the condition that a specific chemical conversion can be performed by a set of given reactions and that this ability will be lost by elimination of any of these reactions. Groups of networks are identified with respect to their ability to perform a certain number of metabolic conversions in an elementary way which are called the network's functions. The number of the network functions defines the degree of multifunctionality. Transitions between networks and mutations of networks are defined by exchanges of single reactions. Different mutations exist such as gain or loss of function mutations and neutral mutations. Based on these mutations neighbourhood relations between networks are established which are described in a graph theoretical way. Basic properties of these graphs are determined such as diameter, connectedness, distance distribution of pairs of vertices. A concept is developed to quantify the robustness of networks against changes in their stoichiometry where we distinguish between strong and weak robustness. Evolutionary algorithms are applied to study the development of network populations under constant and time dependent environmental conditions. It is shown that the populations evolve toward clusters of networks performing a common function and which are closely neighboured. Under changing environmental conditions multifunctional networks prove to be optimal and will be selected.

With these two models completely novel insight is gained regarding both the biological aspect of the structural design of metabolic systems as well as the methodology on which the performed investigations are based. In chapter 4 several possible future applications are proposed and – especially in the field of regulatory mechanisms – first results are presented and guidelines for the development of the appropriate model extensions are given. In general, the methods developed in this work show a wide applicability for the structural analysis of related biological and biophysical systems.

Zusammenfassung

Diese Arbeit basiert auf der grundlegenden Annahme, dass sämtliche Stoffwechselsysteme, so wie sie in heutzutage lebenden Organismen vorzufinden sind, als ein Ergebnis eines evolutionären Prozesses basierend auf Mutationen und natürlicher Selektion betrachtet werden können. Dies betrifft insbesondere die kinetischen Parameter der Enzyme als auch die Stöchiometrie des Systems. Es kann daher angenommen werden, dass diese Parameter im Laufe der Evolution Werte angenommen haben, die gewisse optimale Eigenschaften bezüglich ihrer biologischen Funktionen wiederspiegeln.

Mit dieser Hypothese als Grundlage wird das strukturelle Design von metabolischen Reaktionssystemen untersucht. Zwei diesbezügliche Modelle werden vorgestellt, die das gemeinsame Ziel verfolgen, strukturelle Eigenschaften von Stoffwechselsystemen mit Hilfe von Optimierunsprinzipien zu erklären.

Das erste Modell, welches in Kapitel 2 behandelt wird, untersucht das strukturelle Design von ATP- und NADH-produzierenden Systemen wie die Glykolyse und den Zitratzyklus. Es wird angenommen, dass diese Stoffwechselwege zusammen mit der oxidativen Phosphorylierung im Laufe ihrer Evolution eine hohe Effizienz bezüglich der ATP-Produktionsraten erreicht haben. Auf der Grundlage von kinetischen und thermodynamischen Prinzipien können Aussagen über die optimale Stöchiometrie solcher Systeme getroffen werden. Als Erweiterung früherer Untersuchungen werden die Konzentrationen sowohl der Adeninnukleotide als auch der Nikotinamidadenindinukleotide als variable Größen angesehen. Dies beinhaltet, dass ein ATP- und NADH-produzierendes System in Wechselwirkung mit einem ATP-verbrauchenden, einem NADH-Verbrauch mit ATP-Produktion koppelnden und einem von ATP-Produktion entkoppelten, NADH verbrauchenden System gesehen werden muss. Durch den Vergleich mit einer großen Anzahl alternativer Wege wird untersucht, inwiefern reale Stoffwechselsysteme als optimal angesehen werden können. Die Kinetiken der einzelnen Reaktionen werden als lineare oder bilineare Funktionen der Metabolitkonzentrationen beschrieben. So kann die ATP-Produktionsrate eines jeden erdenklichen ATP- und NADH-produzierenden Weges errechnet werden. Es stellt sich heraus, dass die meisten möglichen Wege eine niedrige ATP-Produktionsrate aufweisen und dass die effizientesten Wege einige strukturelle Gemeinsamkeiten besitzen. Die Optimierung bezüglich der ATP-Produktionsrate wird mit einem evolutionären Algorithmus durchgeführt. Folgende Resultate stehen mit dem tatsächlichen Design der Glykolyse und des Zitratzyklus in Einklang: (i) In allen effizienten Wegen befinden sich die ATP-verbrauchenden Reaktionen am Anfang. (ii) In allen effizienten Wegen befinden sich die sowohl die NADH- als auch die ATP-produzierenden Reaktionen am Ende.

(iii) Die Anzahl der NADH-Moleküle, die aus einem energiereichen Molekül (Glukose) produziert werden, beläuft sich in allen effizienten Wegen auf vier. Um Unterschiede und Ähnlichkeiten verschiedener Wege zu analysieren, wurde ein Distanzmaß und ein Maß für die interne Anordung der Reaktionen eingeführt.

Das zweite Modell, welches in Kapitel 3 vorgestellt wird, folgt einem allgemeineren Ansatz. Ausgehend von einer geringen Anzahl an Reaktionen, die Änderungen des Kohlenstoffskeletts der beteiligten Metabolite beschreiben, werden vollständige Mengen metabolischer Netzwerke konstruiert. Diese werden durch die Anzahl und den Typ der beteiligten Reaktionen charakterisiert. Elementare Netzwerke werden dadurch definiert, dass eine bestimmte chemische Umwandlung durchgeführt werden kann und dass diese Fähigkeit verloren geht, wenn eine der beteiligten Reaktionen ausgeschlossen wird. Netzwerke werden bezüglich ihrer Fähigkeit, eine bestimmte Anzahl von Umwandlungen elementar durchzuführen, gruppiert. Solche Umwandlungen werden Funktionen eines Netzwerkes genannt. Die Anzahl der Funktionen eines Netzwerkes bestimmt dessen Grad an Multifunktionalität. Übergänge zwischen Netzwerken und Mutationen werden durch den Austausch einer einzigen Reaktion definiert. Es existieren verschiedene Mutationen, solche bei denen Funktionen verloren gehen, welche dazugewonnen werden, und neutrale Mutationen. Mutationen definieren Nachbarschaftsrelationen, die graphentheoretisch beschrieben werden. Eigenschaften wie Durchmesser, Konnektivität und die Abstandsverteilung der Vertizes werden berechnet. Ein Konzept zur Quantifizierung der Robustheit von Netzwerken gegenüber stöchiometrischen Veränderungen wird entwickelt, wobei zwischen starker und schwacher Robustheit unterschieden wird. Evolutionäre Algorithmen werden angewandt, um die Entwicklung von Netzwerkpopulationen unter konstanten und zeitlich veränderlichen Umweltbedingungen zu untersuchen. Es wird gezeigt, dass Populationen sich zu Gruppierungen von Netzwerken hinentwickeln, die gemeinsame Funktionen besitzen und nah benachbart sind. Unter zeitlich veränderlichen Umweltbedingungen zeigt sich, dass multifunktionelle Netzwerke optimal sind und sich im Selektionsprozess durchsetzen.

Mit diesen zwei Modellen werden vollständig neue Erkenntnisse gewonnen sowohl bezüglich des biologischen Aspektes des strukturellen Designs von Stoffwechselsystemen als auch bezüglich der Methoden, mit deren Hilfe die hier vorgestellten Untersuchungen durchgeführt werden. In Kapitel 4 werden mögliche zukünftige Anwendungen vorgeschlagen und – besonders in Bezug auf regulatorische Mechanismen – erste Resultate präsentiert. Insgesamt zeigen die hier entwickelten Methoden eine breite Anwendbarkeit auf die strukturelle Untersuchung biologischer und biophysikalischer Systeme.

Contents

Chapter 1

Introduction

1.1 Historical overview

In recent decades experimentalists have worked hard and successfully to unravel the mechanisms of the metabolism of cellular organisms. Nowadays, we have a vast knowledge on these systems, regarding both the detailed viewpoints such as enzymatic activities and regulatory mechanisms as well as the complex structure, the "topology" of metabolic systems. One look in the biochemical atlas (Michal 1999) which summarises our knowledge on the design of metabolic pathways very vividly, tells us how immense the diversity of biochemical reaction systems is.

Now that this accumulation of facts is in a far advanced state and is likely to be completed in the near future, almost necessarily a phase of theoretical analysis follows with the purpose to understand how this mosaic of innumerable fragments of knowledge can be put together in a general context to reveal the overall picture.

Not surprisingly, metabolic networks are among the most thoroughly studied biological systems in the field of mathematical modelling. The classical approach of modelling metabolic systems is mainly concerned with the simulation of the time-dependent behaviour of the variables at fixed values of the parameters by using systems of ordinary differential equations. This approach dates back to the pioneering work of Garfinkel and Hess (1964) on glycolysis. Later on, this metabolic chain and related pathways of cellular energy metabolism were favoured subjects of successful mathematical modelling, early work mainly concerning the energy metabolism in erythrocytes (see e. g. Rapoport et al. 1976, Werner and Heinrich 1985, Joshi and Palsson 1989, Joshi and Palsson 1990, Mulquiney and Kuchel 1999a, Mulquiney and Kuchel 1999b). More recent models aim to simulate the dynamics of energy metabolism of yeast cells (Rizzi et al. 1997, Wolf

and Heinrich 2000, Teusink et al. 2000, Wolf et al. 2001).

Whenever "parameters" are mentioned, generally well-defined quantities are meant whose values have been measured at some point in the past and as a rule can be considered to be constant in time. In contrast to these parameters, the variables of a model describe quantities which may change considerably on a short time-scale and which determine the short-term behaviour of a system.

All the models mentioned above aim at reproducing observed behaviour and thus gaining insight in the mechanisms that regulate the system. The simulated variables depend on the parameters which have to be provided to the model as input data. Among these parameters are of course the structural properties of metabolic pathways, such as their stoichiometry and the values of the kinetic parameters.

Considering the success of the various models (for an overview see Heinrich and Schuster 1996), the question arises whether not only the variables but also the parameters can be explained by models. It is clear that such models pursuing a completely different aim than the before mentioned kinetic models necessitate an entirely new approach.

It has often been stated that the structural design of existing metabolic pathways in cells may be considered as the optimal outcome of selection processes during evolution (Meléndez-Hevia and Isidoro 1985, Meléndez-Hevia and Torres 1988, Heinrich and Hoffmann 1991, Angulo-Brown et al. 1995, Nuño et al. 1997, Mittenthal et al. 1998, Waddell et al. 1999). It should be stressed at this point that of course there exists no such thing as "the optimised organism". Clearly, different organisms show differences in the designs of their metabolism. In the glycolytic pathways of bacteria, for example, the enzyme hexokinase does not exist. Instead, its function – the phosphorylation of glucose – is performed by the phosphotransferase system (see e. g. Saier Jr. 2002), which essentially transfers a phosphate group from phosphoenolpyruvate to glucose. In other organisms, the role of the molecule ATP as the central molecular energy unit is taken over by GTP and NADH is sometimes replaced by NADPH. However, the general overall design of central metabolic pathways such as glycolysis and – considering only aerobic organisms – the citric acid cycle are almost identical in all extant organisms. Due to the simplifications made in the presented models, minor differences such as those mentioned above do not play a significant role in the present context and therefore the metabolic systems under investigation can indeed be understood as the optimal outcome of an evolutionary process.

By this rationale the "parameters" are indeed not so constant as they seem at first sight when investigating systems on a short time scale. On an evolutionary time

scale, i. e. on time scales orders of magnitudes longer than the life-span of any single organism, these parameters must be considered to be subject to mechanisms such as natural selection driving the evolutionary process.

Therefore, a consequential approach to explain the structural parameters of metabolic systems is to make use of optimisation principles. Early work in this direction concerned optimisation of kinetic properties and concentrations of enzymes (Heinrich et al. 1987) as well as optimal properties of the stoichiometry of the pentose phosphate cycle (Meléndez-Hevia and Torres 1988).

Metabolic optimisation has also been performed by Varma and Palsson (1993) in order to identify metabolic routes which are characterised by a maximal yield of certain metabolic compounds such as ATP, NADH, NADPH etc. per consumed molecule glucose. The analysis is based on the given stoichiometric scheme of the central metabolism of *E. coli*. Kinetic properties of this metabolic system have not been taken into account. Such an analysis is certainly a very important prerequisite of successful optimisation in the framework of metabolic engineering but cannot explain the evolutionary emergence of the special stoichiometric design found in contemporary cells.

1.2 Research objectives

The models presented in this work aim to overcome this restriction. The general principle of the models is that they require very little input information but rather rules defining boundary conditions. In this way it is possible to design a model in such a fashion that – by only providing some general rules and guidelines – it can *by itself* develop designs of metabolic reaction systems. So these designs are actually the *result* of the models instead of a priori input specifications.

A central part of each model is a schematic description of metabolic systems allowing for the representation of theoretically possible, chemically feasible alternatives. This description is based on a set of reactions which are used as building blocks for the alternative systems. The exact choice of the set of these reactions depends strongly on the special case investigated. However, in order to allow for a great variety of alternative systems, the individual reactions are characterised not in very much detail concerning the chemical transformations catalyzed by the corresponding enzymes and their kinetic properties. Instead, only classes of different *generic* reactions are taken into account.

A model with the ability to describe a vast amount of structurally different alternative designs can be used to perform various kinds of analysis, such as finding

the optimal design under certain criteria and comparison of specific properties of the alternatives with the design actually found in contemporary cells.

In chapter 2 this principle is applied to investigate the structural design of the central energy metabolism using optimisation principles. The generic reactions have been chosen to allow for the description of a huge number of chemically feasible alternative linear arrangements of ATP and NADH producing and consuming reactions and an evolutionary algorithm as an efficient optimisation procedure has been applied to identify properties of optimal designs regarding a high ATP output rate.

In chapter 3 the restriction to linear arrangements of reactions is lifted and the description of branched and cyclic network structures is allowed for. Inspired by frequently occurring stoichiometric motifs of enzymatic reactions taking place in cellular metabolism, the model is based on a set of generic reactions splitting or merging carbon containing compounds or transferring groups of carbons from one compound to another. Additionally to a complete analysis with respect to the whole class of possible network structures, some evolutionary scenarios are presented in detail.

1.3 Evolutionary algorithms

In the present work, evolutionary algorithms are heavily used both as an efficient optimisation tool and as a method to model a possible evolutionary behaviour of a population of cellular organisms. Due to their methodological importance this whole section shall be dedicated to these algorithms.

Consider the following problem: Out of a huge – possibly infinite – set $\mathcal{S}$ (search space) of entities (individuals) the "best" shall be selected. This necessitates on the one hand a method to evaluate the single individuals, i. e. a mathematical function $\phi : \mathcal{S} \to \mathbb{R}$ is needed which assigns each entity a number measuring its quality. In the following it shall be assumed that a high value of ϕ indicates a higher quality (whether the function ϕ is maximised or minimised is technically the same problem). On the other hand a search strategy is needed to search through the search space $\mathcal{S}$. For small sets $\mathcal{S}$ with a limited number of elements, this search can be performed systematically. However, for larger or even infinite sets the systematical approach obviously provides no option. A good alternative is the application of an evolutionary algorithm providing an intelligent search strategy in large sets (Rechenberg 1989; Goldberg 1989).

The functioning of the algorithm used in this work is essentially the following:

1. In the initialisation step, a certain number N_{pop} of individuals (elements of $\mathcal{S}$)

are randomly created. This necessitates the ability (subroutines, or "methods" in the language of the object-oriented programmer) to create a random element of $\mathcal{S}$.

2. The mutation step applies mutation and crossover operations with some probabilities p_{mut} and p_{cross}, respectively, thus altering some of the individuals. For this step, the corresponding methods have to be provided. Mutation and crossover rules must be carefully designed in order to guarantee firstly that the resulting individual after application of the corresponding operator is still an element of $\mathcal{S}$, and secondly that in principle all elements of $\mathcal{S}$ can be reached by a finite number of such operations. Clearly, if the latter condition remains unfulfilled, there exists always a subset of $\mathcal{S}$ which has not been searched and therefore the result of the algorithm is unreliable. In principle, one of the operations – mutation or crossover – suffices for the algorithm to work. It may, however, alter its efficiency and its reliability (see below). In practical use, very good results have been obtained without the use of crossover mechanisms.

3. The evaluation step assigns each element of the population its corresponding "fitness" value ϕ. Therefore, the function ϕ is also called the fitness function.

4. In the replication step, every element is duplicated with a probability r that correlates monotonously to the fitness value determined in the previous step.

5. Finally, some randomly selected elements of the population are eliminated to reduce the size of the population to its initial value N_{pop}. This step simulates a selection pressure.

Steps 2–5 are repeated until an exit condition is met that has to be specified. For obvious reasons, a loop through the steps 2–5 is called a generation. In most cases it is sufficient to observe the algorithm several times and chose a fixed number of generations after which the procedure stops. This number should be selected in such a way that the fitness of the best elements of the population remains constant over a considerable number of generations. In this case it can be assumed that an optimum has been reached.

During the process, it is regularly observed that a population consists of many copies of the individual yielding the highest fitness value ϕ within the population (the master species) and such individuals resulting from the master species by a small number of mutations. As soon as an individual with a higher fitness value than the

previous master species is found, this individual will reproduce itself with a higher probability and therefore possibly act as a seed for establishing a new master species. The population can be picturised as a "cloud" moving through the search space while, due to the selective pressure, it tends to travel towards regions with a higher fitness value. At an optimal state, the cloud remains more or less constant in its shape. This state is also denoted as "quasi-species" (Eigen 1971; Eigen et al. 1989).

With this intelligent search strategy only a small subset of the whole search space is actually evaluated. Nevertheless, the results generated by this kind of algorithm prove extremely satisfactory. However, it is quite possible for an algorithm to get "stuck" in a local sub-optimum. How likely it is for this to happen, strongly depends on the structure of the "fitness-landscape", which can be envisaged as the values of the fitness function ϕ plotted over the whole search space $\mathcal{S}$. It is especially difficult to escape sub-optima if a large number of mutation steps is needed to reach regions with a higher fitness value. To a certain extent, this problem can be overcome by the introduction of crossover mechanisms as are mentioned in step 2 of the algorithm. In general, since this optimisation procedure does not locate the global maximum with absolute certainty, the procedure has to be repeated several times, the process has to be observed and parameters such as the mutation and crossover probabilities have to be adjusted.

In section 2.2 the application of an evolutionary algorithm is described which is used as an optimisation method to determine the most efficient ATP producing reaction sequences. However, evolutionary algorithms provide a tool not only for determining optima of a given function, but they can also be utilised to simulate evolutionary processes. If, for example, one is interested in how a population of individuals behaves when external conditions change, the same algorithm as described above can be used with just some minor changes. If the fitness function ϕ is not considered to be constant with time but rather varies to reflect changes in the environment, the development of the population can be observed. This provides an interesting simulation tool for time-changing scenarios. In section 3.6 an evolutionary algorithm has been used to observe changes in a population of metabolic networks with different biological functions under changing environmental conditions. Using a fitness function ϕ which is altered not only by externally applied but nevertheless fixed rules, but rather one which is influenced by the individuals of the population themselves, even the interaction of individuals can be simulated. An example to this approach is presented in section 3.6.2 where the interaction of individuals of a population of reaction networks with different metabolic functions is simulated by the fact that some reaction networks can metabolise the products of others.

Chapter 2

Unbranched reaction systems producing ATP and NADH

In this chapter a model is presented which is used to examine the hypothesis that the ATP production rate was an important target at the evolutionary optimisation of cellular energy metabolism corresponding to the main biological function of the underlying pathways (cf. Ferea et al. 1999). In extension to previous work (Heinrich et al. 1997; Meléndez-Hevia et al. 1997; Stephani and Heinrich 1998; Stephani et al. 1999) the present model takes into account not only anaerobic ATP production taking place in glycolysis but also aerobic ATP production via production of NADH and subsequent oxidative phosphorylation.

In order to detect pathways with an optimal structure a vast number of stoichiometrically possible reaction chains are constructed using different types of reactions and certain combinatorial rules. On the basis of an appropriate kinetic description of the individual reactions and subsequent calculation of the steady state ATP production rate optimal reaction sequences are selected. For an efficient optimisation a genetic algorithm (see section 1.3) is applied. Starting from a population of random sequences, this algorithm creates new feasible pathways by applying certain mutation rules and selects in every subsequent generation those pathways characterised by high values of the performance function. The goal of this strategy is to identify those pathways characterised by maximal ATP production rate. The mutation rules have been defined in such a way that all possible sequences can theoretically be the result of a finite number of mutations applied to any other reaction sequence. Attention is paid to the close interrelation of a stoichiometric and kinetic description of metabolic systems. To allow for a great variety of alternative systems, the individual reactions are not characterised

in detail. Instead, classes of different *generic* reactions are taken into account and all reactions are described by simple linear or bilinear kinetic equations.

It is shown that despite the simplicity of the model assumptions the evolutionary optimisation strategy leads to pathways which show many stoichiometric features which are also found in cellular energy metabolism, in particular the ATP and NADH producing system of glycolysis combined with the TCA cycle.

2.1 The model

2.1.1 Stoichiometric properties

We confine our analysis of the evolutionary optimisation of biochemical networks to unbranched pathways. On the basis of various stoichiometric rules different chemically feasible pathways are constructed which are compared with respect to a performance function as is explained below. A pathway C involves r_C reactions transforming an initial substrate X_0 into an end product X_{r_C} via r_C-1 intermediates X_i^C. The reactions allowed are called *generic* since they represent certain *types* of reactions rather than special biochemical reactions specified by the chemical nature of all their substrates and products. These reactions act on molecules consisting of a *skeleton* S_j having two binding sites. Each binding site may be occupied by a ligand which is in the following either a hydrogen atom (H) or a phosphate group (P). For each S_j there exist, therefore, 9 different states $S_j^{(k)}$ as shown in Fig. 2.1 (the symbol 0 denotes an empty binding site of S_j). We neglect the details of the internal structures of the skeletons and assume firstly that each $S_j^{(k)}$ may be transformed into a molecule $S_{j+1}^{(k)}$ by a generic reaction 'u' leading to another skeleton structure but leaving the ligand state unchanged. Here, $j \in \{0,\ldots,U\}$ is an index increasing by one for each 'u'-reaction, and U is the total number of 'u'-reactions in the pathway. Secondly, phosphorylation (dephosphorylation) as well as protonation (deprotonation) may connect different states $S_j^{(k)}$ and $S_j^{(k')}$. Accordingly, for a given pathway C the set of metabolic intermediates consists of subsets of all possible states $S_j^{(k)}$ with $j \in \{0,\ldots,U\}$ and $k \in \{1,\ldots,9\}$.

A full list of generic reactions is given in Table 2.1, which shows that phosphorylations and dephosphorylations may take place not only by a direct uptake or dissociation of phosphate groups (reactions P and p, respectively) but also under participation of the adenine nucleotides ADP and ATP (reactions A and a, respectively). In a similar way we distinguish a direct uptake or removal of hydrogen atoms (reactions H and h, respectively) from reactions which occur with the cofactors NADH and NAD as

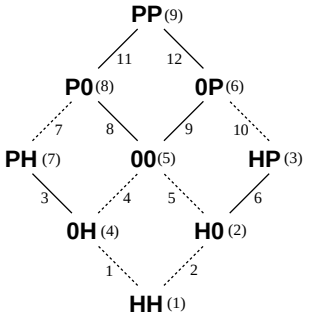


Figure 2.1: Schematic display of the nine different ligand states and the reactions allowed. Dashed lines represent reactions involving hydrogen (h or n in upward direction, H or N in downward direction), solid lines represent reactions changing the phosphorylation state (P or A in upward direction, p or a in downward direction). The letters describe the ligands bound to the ligand sites, the numbers in brackets denote the index of the corresponding ligand state.

additional substrates or products (reactions N and n, respectively).

We do not assume that reactions H and h represent separate enzymatic reactions. In contrast, we will interpret each of them always in combination with another chemical reaction. For example, the phosphorylation of glucose to form glucose-6-phosphate by consuming ATP will in the present model be described as the combination of the two reactions 'h' and 'A'. A separate consideration of the reactions h and H allows for much simpler rules for constructing pathways of different stoichiometry. In order to avoid that this procedure affects the overall kinetics of these pathways, the reactions h and H are considered to be very fast compared to the other reactions (see below).

Please note that any reaction described by an upper case letter denotes the reverse of the reaction described by the corresponding lower case letter. All reactions changing only the ligand state of a certain skeleton (P, p, A, a, H, h, N and n) belong to the subclass of *coupling reactions* whereas the 'u' reactions are called *uncoupled reactions*.

Chemically feasible alternative metabolic pathways are generated by assembling generic reactions fulfilling (a) the boundary condition that the first metabolite X_0 and the last metabolite X_{r_C} have to be in the ground state ($S_0^{(1)}$ and $S_U^{(1)}$, denoted by `HH` in Fig. 2.1), and (b) that the pathway is interconnected (the substrate of any reaction is the product of the preceding reaction). We select `HH` to be the starting point, because this roughly corresponds to the non-oxidized state of the glucose molecule, with which the state $S_0^{(1)}$ is identified. Furthermore, we exclude cases where a metabolic intermediate appears more than once in a pathway. All reactions are considered reversible.

Symbol	Generic Reaction
a	Transfer of one phosphate group from a metabolic intermediate to ADP, resulting in the production of one molecule ATP
A	Transfer of one phosphate group from ATP to a metabolic intermediate, resulting in the consumption of one molecule ATP
p	Release of one (inorganic) phosphate group from a metabolic intermediate
P	Binding of one (inorganic) phosphate group to a metabolic intermediate
n	Transfer of one hydrogen atom from a metabolic intermediate to NAD, resulting in the production of one molecule NADH
N	Transfer of one hydrogen atom from NADH to a metabolic intermediate, resulting in the consumption of one molecule NADH
h	Release of one hydrogen from a metabolic intermediate
H	Binding of one hydrogen to a metabolic intermediate
u	uncoupled reaction

Table 2.1: List of generic reactions

We define the "forward" direction by calling $S_0^{(1)}$ the initial substrate and $S_U^{(1)}$ the final product.

Based upon Fig. 2.1 we introduce a graphical representation of the pathways (see Fig. 2.2 for an example). Any coupling reaction is represented by a solid arrow connecting two ligand states along the edges of the graph shown in Fig. 2.1. Any uncoupled reaction is represented by a long dashed arrow connecting two skeletons and starting and ending in the same ligand state.

A solid arrow specifies one of the reaction pairs (H,N), (h,n), (P,A) or (p,a) by its location and direction within the graph of the ligand states (Fig. 2.1). Arrows standing for phosphorylations point upwards, arrows standing for hydrogen uptake point downwards. However, any solid arrow may stand for one of two different generic reactions from the pairs mentioned above. Neglecting the nature of these coupled reactions yields an arrangement of arrows which we call the *topology* of the pathway. Hence, a unique description of a given pathway is obtained by adding the given types

of the coupled reactions to the corresponding arrows (see Fig. 2.2).

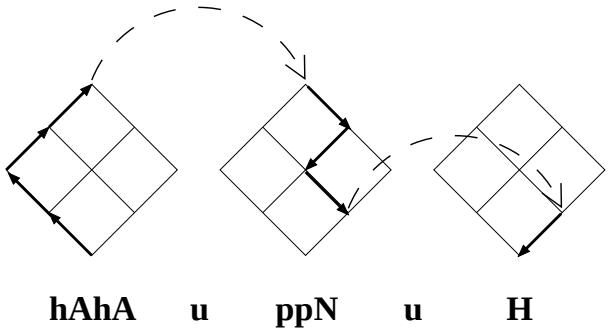


Figure 2.2: Graphical notation of the pathway **hAhAuppNuH**

Obviously, these special rules for the construction of simple unbranched chains allow to consider glycolysis as a special case. There, maximally two phosphate groups may be bound to the carbohydrates at sites which may be occupied also by hydrogen atoms. Furthermore, there exist ATP consuming reactions (hexokinase and phosphofructokinase), ATP producing reactions (phosphoglycerate kinase and pyruvate kinase), direct phosphorylations (glyceraldehyde dehydrogenase) and splitting of phosphate groups (bisphosphoglycerate phosphatase). In glycolysis, uptake and release of hydrogen may take place directly or under participation of the NAD/NADH system (catalyzed by glyceraldehyde dehydrogenase, and, under anaerobic conditions, catalyzed by lactate dehydrogenase or alcohol dehydrogenase).

Moreover, the stoichiometric rules of our model allow to construct, to a certain degree, the reactions of the citric acid cycle (TCA) which are characterised by a net production of NADH.

Compared with the real situation, the above stoichiometric description contains some crude simplifications which have to be kept in mind. In glycolysis, a branching occurs at the aldolase reaction where fructose-1, 6-bisphosphate (with six carbon atoms) is split into glyceraldehyde-3-phosphate and dehydroxyacetone phosphate which both contain three carbon atoms. These two isomers are interconverted into each other by triose phosphate isomerase. This branching is neglected in the present model. Secondly, concerning the citric acid cycle the stoichiometric rules of our model make it possible to describe NADH production (catalyzed in TCA by the enzymes isocitrate dehydrogenase, α-ketoglutarate dehydrogenase, and malate dehydrogenase) but it is evident that an unbranched sequence is not sufficient to account for a cyclic arrange-

ment of reactions. However, also in TCA a high number of subsequent reactions are arranged in a linear sequence (the 8 reactions transforming citrate into oxaloacetate) and what we neglect is the stoichiometric and kinetic feedback occurring at the reaction catalyzed by citrate synthase.

Concerning the main goal of our analysis, that is the evolutionary optimisation of ATP producing pathways, we evaluate different possible reaction sequences C (assembled using the rules explained above) by comparing their ATP production rates under steady state conditions. Obviously, this necessitates not only the consideration of a pathway C but also the incorporation of external ATP consuming processes (ATPases) as well as external NADH consumption. With respect to the latter case we take into account a non-specific NADH consumption (d) as well as oxidative phosphorylation (Ox). The latter process also accounts for ATP production which is external with respect to the main reaction chain C.

Restricting the number of ligands to two also restricts the number of hydrogen atoms allowed to take part in the reactions. However, there are twelve hydrogen atoms bound to glucose. The model assumptions can be justified by classifying the reactions involving hydrogen into two classes. The first class consists of reactions that are functionally related to NADH production / consumption or to a change in state of phosphorylation. The second class consists of reactions like hydrolysis. These reactions are not directly related to the principal functionality of the pathway, i. e. ATP / NADH production, so we include these types of reaction in the 'uncoupled' reactions.

2.1.2 Kinetic properties

A calculation of the ATP production rate necessitates the consideration of the kinetic properties of the main components of our model, i. e. the reaction sequence C, the external ATPases, oxidative phosphorylation and the non-specific NADH consumption.

The first and last metabolites in any unbranched chain C are considered external and their concentrations are kept fixed. The concentrations of adenine nucleotides as well as nicotinamide adenine dinucleotides are considered variable quantities with the restriction that the sum of ATP and ADP as well as the sum of NADH and NAD are constant. We denote the concentrations of ADP and ATP by A_2 and A_3, respectively, the concentrations of NAD and NADH by N_1 and N_2, respectively. The overall concentrations of adenine nucleotides and nicotinamide adenine dinucleotides

are denoted by A and N, respectively. We introduce the relative concentrations

$$a_2 = \frac{A_2}{A}, \qquad a_3 = \frac{A_3}{A}, \tag{2.1}$$

$$n_1 = \frac{N_1}{N}, \qquad n_2 = \frac{N_2}{N}, \tag{2.2}$$

which implies $a_2 = 1 - a_3$ and $n_1 = 1 - n_2$. Obviously

$$0 \leq a_2, a_3, n_1, n_2 \leq 1. \tag{2.3}$$

For the bimolecular reactions 'A' / 'a' and 'N' / 'n' bilinear kinetic equations are used. The 'u'-reactions are described by linear kinetic equations. Concerning the reactions H, h, P and p we assume that the concentrations of protons as well as inorganic phosphate are constant. In this way these reactions are described by pseudo first-order kinetic equations. The kinetic parameters are assumed to be the same for all reactions within a given class.

For given values of a_3 and n_2 the rate equations of all reactions are linear in the concentrations of the metabolites X_i. Thus, quasi monomolecular rate equations can be used and the following expression for the steady-state rate J_C holds true (see Heinrich and Schuster 1996)

$$J_C = \frac{X_0 \prod_{j=1}^{r_C} q_j - X_{r_C}}{\sum_{j=1}^{r_C} \frac{\tau_j(1+q_j)}{q_j} \prod_{k=j}^{r_C} q_k}. \tag{2.4}$$

Here, q_j and τ_j denote the equilibrium constants and relaxation times for the generic reaction they describe.

For bimolecular reactions (all reactions involving either ADP / ATP or NAD / NADH, i. e. A, a, N and n), q_j and τ_j are *effective* quantities depending on the concentrations of the cofactors. Let us consider for example an ATP consuming process

$$X_i + ATP \rightleftharpoons X_{i+1} + ADP \tag{2.5}$$

characterised by the rate equation

$$\nu_A = \kappa_+^A \cdot X_i \cdot A_3 - \kappa_-^A \cdot X_{i+1} \cdot A_2, \tag{2.6}$$

with the second order rate constants κ_+^A and κ_-^A. An analogous equation can be derived for ATP producing reactions by keeping in mind, that $\nu_a = -\nu_A$. With the relaxation time

$$\tau_A = \frac{1}{\kappa_+^A \cdot A_3 + \kappa_-^A \cdot A_2} \tag{2.7}$$

and the equilibrium constant

$$q_A = \frac{\kappa_+^A}{\kappa_-^A}, \tag{2.8}$$

Eq. (2.6) for reaction (2.5) can be rewritten as

$$\nu_A = \frac{\left(q_A \cdot \frac{A_3}{A_2}\right) \cdot X_i - X_{(i+1)}}{\tau_A \left(1 + q_A \cdot \frac{A_3}{A_2}\right)}. \tag{2.9}$$

The expression $q_A \cdot A_3/A_2$ is an effective equilibrium constant and has to be used as q_j in Eq. (2.4) for any 'A'-reaction. Introducing the relaxation time

$$\widetilde{\tau_A} = \frac{2}{A\,(\kappa_+^A + \kappa_-^A)} \tag{2.10}$$

at the reference state $A_2 = A_3 = A/2$, Eq. (2.7) can be written as

$$\tau_A = \widetilde{\tau_A} \cdot \frac{A(1 + q_A)}{2(A_2 + q_A \cdot A_3)}. \tag{2.11}$$

This relaxation time has to be chosen for any τ_j in Eq. (2.4) belonging to an 'A'-reaction.

Analogous expressions have been derived for the reactions a, N and n.

As the effective equilibrium constants q_A and q_N depend on the concentrations a_3 and n_2 respectively, J_C is a function of these two variables.

In order to calculate the ATP production rate of any given chain C of reactions, we consider the interaction of three systems. Firstly, the ATP and NADH producing system is described by a sequence of reactions defining the unbranched chain of generic reactions, as explained in section 2.1.1. Secondly, an external ATP consuming reaction is considered, and thirdly a system that consumes NADH and uses the reducing power in order to produce ATP. The real equivalent of the first system can be thought of as glycolysis and the citric acid cycle, of the second as a representation of all external ATPases, of the third as oxidative phosphorylation. The flux of the first system is described by Eq. (2.4). The rate of the second process is assumed to be

$$J_{ATPase} = k_{ATPase} \cdot A_3 = A \cdot k_{ATPase} \cdot a_3, \tag{2.12}$$

where k_{ATPase} is the rate constant. The stoichiometry of the third system is depicted in Fig. 2.3. As in the real oxidative phosphorylation more than one molecule ATP is produced for each molecule NADH, we introduce the stoichiometric factor γ. In our calculations, we use the realistic value $\gamma = 3$ (see Stryer 1988). The branch with

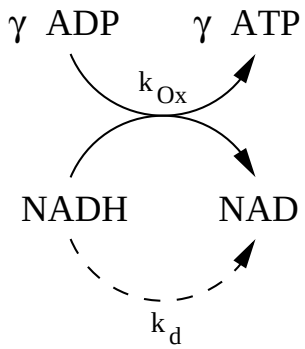


Figure 2.3: Production of ATP and consumption of NADH outside reaction chain C

rate constant k_d has been introduced to account firstly for a consumption of NADH as it occurs in many biosynthetic reactions (mediated by the NADH / NADPH transhydrogenase) and secondly for the decoupling of NADH consumption and oxidative phosphorylation at the inner mitochondrial membrane. Thus the third system can be described by the two equations

$$J_{Ox} = k_{Ox} \cdot A_2 \cdot N_2 = A \cdot N \cdot k_{Ox} \cdot (1 - a_3) \cdot n_2, \tag{2.13}$$

$$J_d = k_d \cdot N_2 = N \cdot k_d \cdot n_2. \tag{2.14}$$

We consider the whole system to be in steady-state and calculate the overall ATP production rate using the balance equations for ATP,

$$d \cdot J_C - J_{ATPase} + \gamma J_{Ox} = 0, \tag{2.15}$$

and for NADH,

$$n \cdot J_C - J_{Ox} - J_d = 0, \tag{2.16}$$

where d and n denote the net production of ATP and NADH molecules, respectively, per consumption of one molecule glucose in reaction chain C. Obviously, d denotes the difference of the numbers of A and a reactions in C, whereas n is the difference of the numbers of N and n reactions.

We eliminate J_C by rewriting Eq. (2.16) as

$$J_C = \frac{J_{Ox} + J_d}{n}. \tag{2.17}$$

Here, we assume that $n > 0$. This means, we explicitely exclude cases with no net NADH production. The case $n = 0$ implies

$$J_{Ox} + J_d = 0, \tag{2.18}$$

and, as J_{Ox} and J_d may attain only non-negative values, this leads to

$$J_{Ox} = J_d = 0. \tag{2.19}$$

Since a vanishing net NADH production makes the part of the model describing oxidative phosphorylation obsolete, we do not consider this case. A model for this case has already been developed by Stephani and Heinrich (1998).

Inserting Eq. (2.17) into Eq. (2.15) yields

$$-J_{ATPase} + \frac{d+\gamma n}{n} J_{Ox} + \frac{d}{n} J_d = 0, \tag{2.20}$$

or, using Eqs. (2.12), (2.13) and (2.14),

$$-A \cdot k_{ATPase} \cdot a_3 + \frac{d+\gamma n}{n} A \cdot N \cdot k_{Ox} \cdot n_2 \cdot (1-a_3) + \frac{d}{n} N \cdot k_d \cdot n_2 = 0, \tag{2.21}$$

which yields

$$a_3 = a_3(n_2) = \frac{\frac{d}{d+\gamma n}\frac{k_d}{A \cdot k_{Ox}} + 1}{\frac{n}{d+\gamma n}\frac{k_{ATPase}}{N \cdot k_{Ox}} + n_2} \cdot n_2. \tag{2.22}$$

Thus, a_3 is a monotonous function of n_2. We can confine the scope of these variables by distinguishing whether or not

$$a_3(n_2 = 1) \leq 1. \tag{2.23}$$

Using Eq. (2.22), the condition (2.23) can be written as

$$\frac{d}{n} \leq \frac{A}{N} \frac{k_{ATPase}}{k_d}. \tag{2.24}$$

The scope of the variables a_3 and n_2 can be summarised as

$$0 \leq n_2 \leq 1 \quad , \quad 0 \leq a_3 \leq \frac{\frac{d}{d+\gamma n} \cdot \frac{k_d}{A \cdot k_{Ox}} + 1}{\frac{n}{d+\gamma n} \cdot \frac{k_{ATPase}}{N \cdot k_{Ox}} + 1} \quad \text{if} \quad \frac{d}{n} \leq \frac{A}{N} \frac{k_{ATPase}}{k_d} \tag{2.25}$$

$$0 \leq n_2 \leq \frac{n}{d} \frac{A}{N} \frac{k_{ATPase}}{k_d} \quad , \quad 0 \leq a_3 \leq 1 \quad \text{if} \quad \frac{d}{n} > \frac{A}{N} \frac{k_{ATPase}}{k_d} \tag{2.26}$$

With the dependency (2.22) the steady-state condition for the ATP production rate becomes

$$d \cdot J_C\left(a_3(n_2), n_2\right) - A \cdot k_{ATPase} \cdot a_3(n_2) + \gamma \cdot A \cdot N \cdot k_{Ox} \cdot (1 - a_3(n_2)) \cdot n_2 = 0. \tag{2.27}$$

After solving this one-dimensional equation for n_2 and determining a_3 by Eq. (2.22), the flux J_C is calculated from Eq. (2.4). The values of the parameters k_{ATPase}, k_{Ox} and k_d have to be estimated. Reasonable estimations for these, as well as for the other system parameters, are given in Appendix A.1.

2.2 Optimisation procedure

With the model described in section 2.1.1 and the kinetic equations derived in section 2.1.2, we developed a method to generate chemically feasible alternative pathways producing ATP and NADH and examine a large number of these with respect to their overall ATP production rate

$$J = d \cdot J_C + \gamma \cdot J_{Ox}. \tag{2.28}$$

The number of theoretically possible pathways C is very large. If we restrict the length of the pathways to a maximal number of 'u'-reactions of six, there are about $5.8 \cdot 10^{24}$ pathways that fulfil the conditions described in section 2.1.1. This number is clearly too large to allow for a systematic examination of all pathways. With every additional 'u'-reaction the number of possible pathways multiplies by about 4500. A derivation of this rule is given in Appendix A.2.

In order to compare real glycolysis to alternative pathways and to examine in what respect the real pathway can be considered optimal, an evolutionary algorithm (see section 1.3) is applied to carry out optimisation calculations. For this purpose a modular designed computer program has been developed. The optimisation strategy applied here follows in principle the strategy described by Stephani et al. (1999). The algorithm is initialised by generating a specified number (N_{pop}) of random pathways. Pathways are denominated by strings of characters, each character defining a generic reaction (see Table 2.1). The subroutine generating random pathways takes into account all stoichiometric restrictions that hold. Mutation and selection are then applied repeatedly. The functionality of the whole algorithm strongly depends on reasonable definitions of mutation and selection.

The selection mechanism is implemented by defining a fitness function that depends on the steady-state ATP production rate any given pathway C yields in combination with oxidative phosphorylation. This means that those pathways yielding a higher production rate have a higher probability of reproduction, i. e. duplicating themselves. After the reproduction the population is reduced to its original number N_{pop} by randomly selecting sequences that are eliminated. These steps lead to a "survival of the fittest" behaviour in a sense, that the more efficient pathways have a higher probability of survival.

A mutation algorithm has been developed that ensures that theoretically any element of the whole sequence space (the space that contains all pathways) can be generated by a finite number of mutations beginning with any other element (see theorem 1

in Appendix A.3). Furthermore, the algorithm works in such a way that the number of changes that occur in a single mutation step is kept as small as possible. The number of characters changed within a string by a mutation is always less than or equal to three. Theoretically, crossover mutations can be defined as well. Two reactions sequences C and D are cut at suitable positions and split into two subchains, C_1C_2 and D_1D_2, respectively. These subchains can be recombined into two new sequences C_1D_2 and C_2D_1. However, we do not consider crossover mutations in this model. The exact definition of the mutation algorithm is given in Appendix A.3.

The efficiency and the quality of this algorithm depends on the way the fitness is calculated from the steady-state ATP production rate. Here, we use the following formula

$$f_i = \frac{J_i/J^{opt}}{1 + \Omega\,(1 - J_i/J^{opt})}, \tag{2.29}$$

where J_i denotes the steady-state ATP production rate of pathway i, J^{opt} denotes the best steady-state ATP production rate of the present population and Ω is an adjustable control parameter. For $\Omega = 0$, Eq. (2.29) reduces to the linear function

$$f_i = \frac{J_i}{J^{opt}} \tag{2.30}$$

and for $\Omega \to \infty$ to

$$f_i = \begin{cases} 0 & \text{for} \quad 0 \leq J_i < J^{opt} \\ 1 & \text{for} \quad J_i = J^{opt} \end{cases} \tag{2.31}$$

The probability of reproduction is then given by f_i (obviously $0 \leq f_i \leq 1$). This means, the larger Ω, the steeper the function f_i and the stronger the selection pressure. During simulations it turned out, that many sequences yield similar output fluxes, therefore we had to choose a high value for Ω. All results presented below have been obtained by using $\Omega = 10^4$.

Another important control parameter is the probability p that mutations occur. It cannot be generally specified how to choose this parameter. The effect of the parameters on the behaviour of the simulations depend very strongly on the specific problem, especially on the behaviour of the fitness function. It turned out that for the probability of the occurrence of mutations a suitable choice is $p = 0.3$.

A general problem that occurs when running genetic algorithms is the reliability. It can never be said with absolute certainty whether a maximum found is actually the global maximum one was looking for. If a population consists only of sequences that resemble pathways at or near a suboptimal local maximum, the escape of this region with a small number of mutations is very unlikely. It is possible that these problems

could be overcome with the introduction of crossover operators, that enable greater changes in the sequences from one generation to the next one. In our simulations we have to decide by comparison with other results whether or not a result seems to resemble a suboptimal state.

2.3 Results

The optimisation procedure described in section 2.2 does not yield exactly one *best* reaction sequence but a rather high number of sequences characterised by very similar ATP production rates. In Table 2.2 the best ten sequences are listed that ever occurred within four simulation runs over 5000 generations with a population of 200 sequences.

Sequence
hAhApNphuAuaAHpHhPupnAAaHaHunnAPauaHHnuPnPaHaH
hAhApNpuNunnHAunupPHaHunnHuPnPaHaH
hAhApNupNunnHAunpuPHaHunnHuPnPaHaH
hAhApNupNunnHAunupPHaHunnHuPnPaHaH
hAhApNupNunnHAnuAuaHaHunnHuPnPaHaH
hAhApNupNunnHAnuAuapPHaHunnHuPnPaHaH
hAhApNupNunnHAnuupPHaHunnHuPnPaHaH
hAhApNupNunnHAnuAauHaHunnHuPnPaHaH
hAhApNupNnunHAnuAuaHaHunnHuPnPaHaH
hAhApNuphAAaNaHunAnaAHpunPAuaaHHnAhuHaHnuPnPaHaH
⋮
nAnPauAaauHAhAaHunpuNHnAhAauPpuNahNH

Table 2.2: The ten best sequences during simulation runs and a very inefficient one. The following parameters were used: $q_u = q_A = q_N = 1000, q_H = 1, q_P = 1/1000, \tau_u = \tau_A = \tau_P = \tau_N = 1, \tau_H = 1/100, k_2 = 2, k_3 = k_4 = 10, \lambda = 0.5, U = 6$. The order of the sequences is presented with decreasing ATP production rate. However, all sequences yield almost the same ATP production rate of $J_{ATP} = 1.8015362401 \pm 10^{-10}$. The inefficient sequence at the bottom of the table yields a production rate of $J_{ATP} \approx 0.0675$.

The parameter values used in these simulations are given in the legend to Table 2.2

(see also Appendix A.1). The relative variations of the resulting ATP production rate J is less than 10^{-10} for all these sequences, indicating that they are almost identical concerning their biological function. However, the chance of randomly generating a reaction sequence yielding such an efficient ATP production rate is almost negligible. Fig. 2.4 shows the fluxes of 10000 randomly generated sequences. More than 80 per

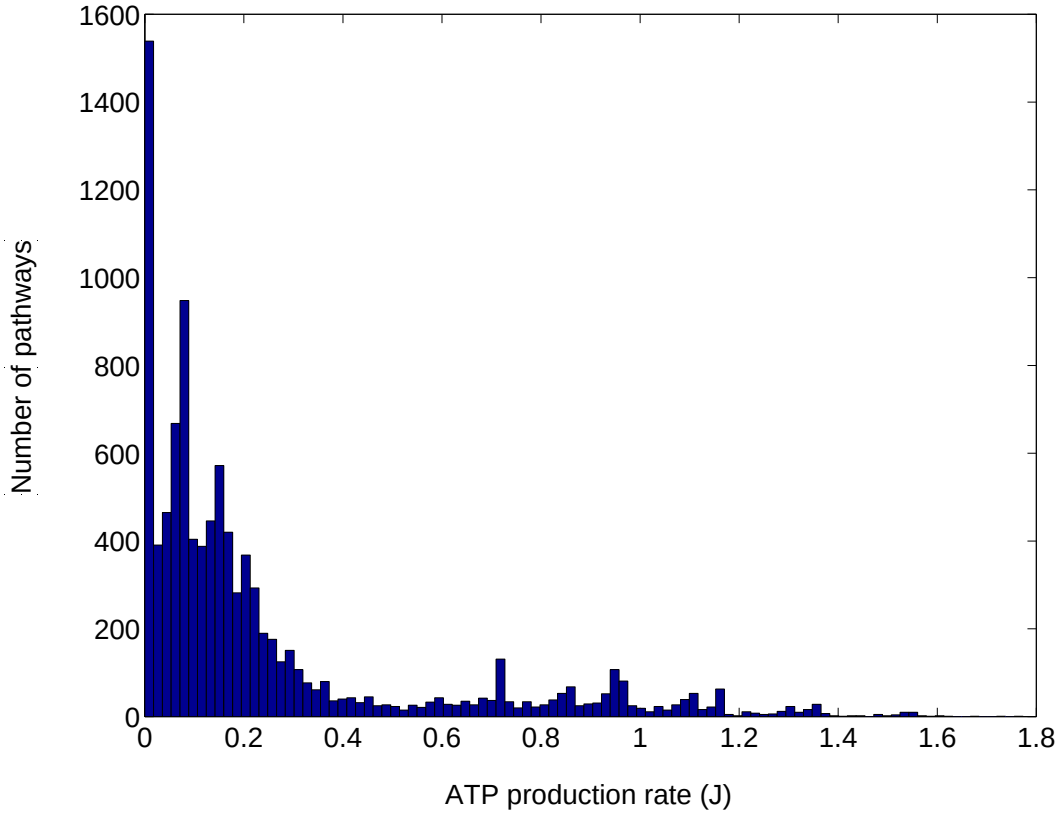


Figure 2.4: Number of pathways out of 10000 randomly generated sequences yielding a given ATP production rate. The fluxes have been calculated with the same parameters as given in the caption of Table 2.2.

cent of the sequences yield a flux smaller than 20 per cent of the optimal flux of the efficient sequences given in Table 2.2. Only five out of the ten thousand yield a flux $J > 1.6$, and only two of these $J > 1.7$. This examination demonstrates the efficiency of the applied evolutionary algorithm.

In the course of the selection process also very inefficient reaction sequences appear. An example for a system with a very low value of J is given in the last row of Table 2.2. All these reaction sequences have the same number of 'u'-reactions ($U = 6$) but differ in

their total length. All optimal sequences are characterised by some common features:

1. For each sequence the number of 'A'-reactions equals the total number of 'a'-reactions, meaning that net ATP production does not take place within the reaction chain C but rather within the subsystem of oxidative phosphorylation. This feature is in accordance with the fact that all chains are characterised by a net production of NADH. For all optimal sequences the difference of the number of NADH producing and NADH consuming reactions is $n = 4$. It is interesting to note that the latter property agrees with the stoichiometry of the citric acid cycle where three molecules NADH and one molecule $FADH_2$ are produced (see e. g. Stryer 1988). Therefore our simulations reproduce an important feature of the real NADH producing pathway, i. e. the net production of four molecules with a reducing power that can be used for ATP production.

2. All optimal reaction chains begin with the subsequence **hAhApN...**. With respect to the first two 'A'-reactions this corresponds to the real design of glycolysis where the ATP consuming reactions catalyzed by hexokinase and phosphofructokinase are also located at the beginning of the pathway.

3. All optimal reaction chains end with the subsequence **...uPnPaHaH**. This characteristic is in line with previous results (see Heinrich et al. 1997; Stephani et al. 1999) that in optimal reaction sequences the ATP producing reactions are located near the end of the pathway – which is in accordance to the location of phosphoglycerate kinase and pyruvate kinase in the lower part of glycolysis.

In contrast to the optimal sequences 1–10, the very inefficient sequence presented in Table 2.2 contains only one ATP consuming reaction at the very beginning and has a lower net production of NADH ($n = 1$).

Unlike in real glycolysis, all optimal sequences 1–10 contain more than two ATP consuming reactions (reaction sequence number 10 even contains 8 such reactions). In order to examine the importance of this feature, we repeated the optimisation procedure with the additional boundary condition that only sequences with at most two ATP consuming reactions are accepted. The ten best sequences out of these simulations are shown in Table 2.3.

Interestingly, the ATP production rates of these sequences are only marginally smaller than those given in Table 2.2. We therefore conclude that limiting the total number of ATP consuming reactions to two is not significantly influencing the outcome of the simulations concerning the biological function of the optimised sequences. A

Sequence
hAhApNuhpuPHuhpHuHnnPHupHnnuHPnPaHaH
hAhApNpuuHnunHHununuHPnPaHaH
hAhApNpuHununHHununuHPnPaHaH
hAhApNupHhunuHPunpHHunnHunPPaHaH
hAhApNupnuuHHnunuHHnuPnPaHaH
hAhApNpuHnuunHuHnunuHPnPaHaH
hAhApNpuHnuunHHununuHPnPaHaH
hAhApNupnuHuuHnnuPHpHnuPnPaHaH
hAhApNuhpNHnunPuHpnNHunnHunHuPnPaHaH
hAhApNupHhhPupPHpHunPunpHHnunHunPPaHaH

Table 2.3: The ten best sequences resulting from simulation runs performed under the condition that the number of ATP consuming reactions must not exceed two. The parameter values are the same as given in the caption of Table 2.2. Moreover, the ATP production rate lies in the same range.

comparison of Tables 2.2 and 2.3 shows that the above mentioned features of optimal sequences are also visible in the subclass of chains with a low number of ATP consuming and ATP producing reactions.

It is interesting to compare the optimised sequences in more detail with the real glycolytic pathway. The 'u'-reactions found in real pathways occur in great diversity, from energetically favourable as an oxidation of an aldehyde group (as occurring in glyceraldehyde-3-phosphate dehydrogenase reaction, see e. g. Stryer 1988) to even slightly unfavourable as phosphoglucose isomerase with an equilibrium constant of about $q = 0.3$ (see Florkin and Stotz 1969). However, as all these reactions are comprised into one generic reaction u, one cannot expect the model to predict the positioning of these reactions. Therefore we do not consider them in the following comparison. Further, we ignore single reactions of type H and h because of the reasons mentioned in section 2.1.1. Combinations **hA** and **aH** are therefore treated as single (de)phosphorylation reactions $\mathbf{A}^\star$ and $\mathbf{a}^\star$, respectively. The upper part of glycolysis (conversion of glucose into fructose bisphosphate) contains two ATP driven phosphorylations and may therefore represented as $\mathbf{A}^\star\mathbf{A}^\star$. The lower part (conversion of glyceraldehyde-3-phosphate into pyruvate) is represented as $\mathbf{nPa}^\star\mathbf{a}^\star$. For comparison of the latter sequence one has to take into account that the glyceraldehyde-3-phosphate dehydrogenase complex can be considered as a combination of three reaction steps, a

phosphorylation (P), an NADH production (n) and a further reaction step (u) which is the above mentioned oxidation of the aldehyde group. Furthermore, the subsequence $\mathbf{a}^\star\mathbf{a}^\star$ is considered to reflect the existence of the two ATP producing steps catalyzed by the enzymes phosphoglycerate kinase and pyruvate kinase in the lower part of glycolysis. We conclude that all optimised sequences represent remarkably well the starting sequence as well as the final sequence of reactions in the glycolytic pathway.

In order to analyse common features and differences of reaction sequences in more quantitative terms, we introduce a distance measure D ($D : \mathcal{S} \times \mathcal{S} \mapsto [0, \infty)$, with $\mathcal{S}$ denoting the sequence space), that enables us to compare the stoichiometries of two reaction sequences C_1 and C_2. The distance measure fulfils the conditions

$$D(C_1, C_2) \geq 0, \text{ equality if and only if } C_1 = C_2 \tag{2.32}$$

$$D(C_1, C_2) = D(C_2, C_1) \tag{2.33}$$

$$D(C_1, C_3) \leq D(C_1, C_2) + D(C_2, C_3) \tag{2.34}$$

For a detailed definition of the distance measure see Appendix A.4.1.

By using the distance measure D, we first compare a set of efficient sequences, all yielding a steady state ATP production rate close to the optimum, with a set of random sequences by calculating the distance between any two sequences. The result is shown in Fig. 2.5 in matrix form. Here, dark spots denote a close resemblance of two sequences, whereas light spots indicate a great difference between two sequences. The sequences numbered by 1–50 are the most efficient sequences that occurred during simulations with the parameter values given in the caption of Table 2.2 (The sequences labelled by 1–10 are the first ten sequences from this table), the sequences numbered by 51–100 have been randomly selected and numbered. We see, that the distances between efficient and random sequences tend to be larger than the distances between random sequences (the off-diagonal quarters of Fig. 2.5 are "lighter" than the upper-right quarter). In the group of sequences with a high ATP production rate, not all pathways are similar to each other. Instead, a structuring in the distances between favourable sequences can be seen. The very dark areas in the lower-left corner of Fig. 2.5 indicate clusters of sequences lying very close together. Fig. 2.5 indicates that sequences 1 and 10 show rather apparent stoichiometric differences with respect to the other optimal sequences 2–9, which is in agreement with the fact that these two sequences are much longer (see Table 2.2). From Fig. 2.5, the important feature of the sequence space can be deduced that there seem to exist several local (sub-)optima.

In order to find a more concise way of identifying clusters of similar sequences, we performed a classical multi-dimensional scaling analysis (see Venables and Ripley 1998),

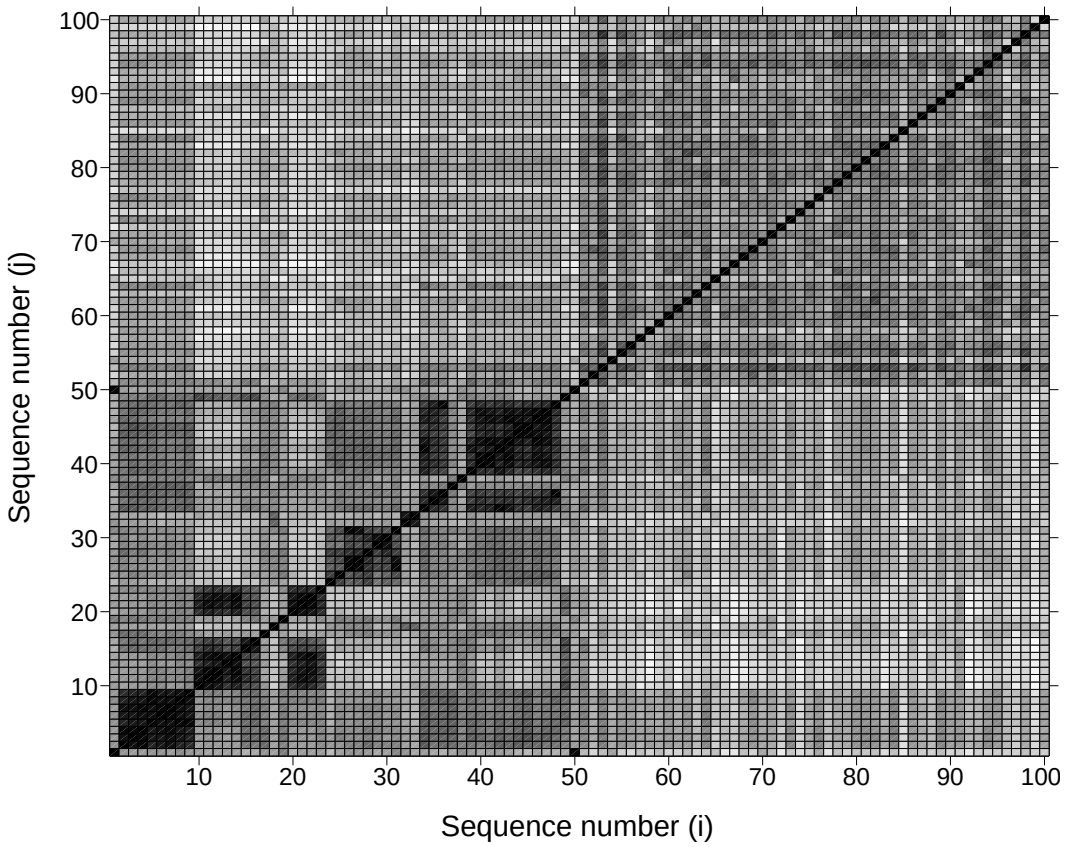


Figure 2.5: Distances $D(C_i, C_j)$ between sequences C_i and C_j. The first 50 sequences are efficient sequences, the other 50 are randomly selected. Dark spots indicate small distances, light spots large distances.

using the entries of the distance matrix underlying Fig. 2.5 as input parameters. In this analysis a two-dimensional set is calculated in which the Euclidean distances represent a best fit to the input data (i. e. the distances between the sequences introduced in Eqs. (2.32)–(2.34) and Appendix A.4.1). The result is shown in Fig. 2.6. We can clearly see that the random sequences (+) are separated by the efficient sequences (∘). The random sequences form a cluster near the top margin of Fig. 2.6, whereas the efficient sequences are spread out along the bottom margin. Within the latter group of sequences, Fig. 2.6 gives few hints of further clustering. Therefore we analysed the same data without the random sequences to detect more structural properties from the distance matrix. The result is shown in Fig. 2.7.

For example, let us take a closer look at the cluster near the centre bottom of

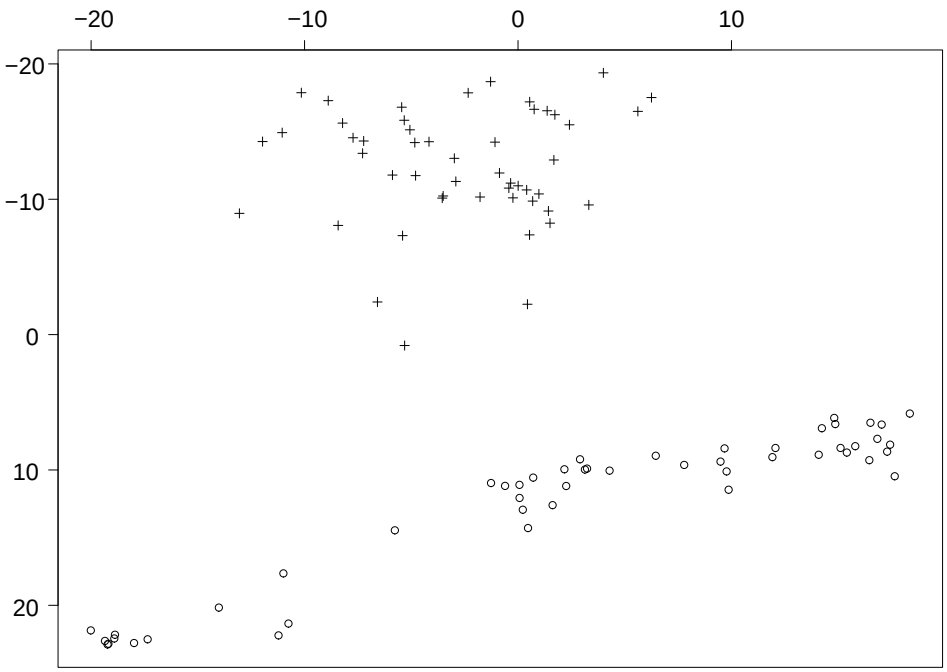


Figure 2.6: Multi-dimensional scaling of the distance matrix in arbitrary units. The random sequences 51–100 are marked by "+", the efficient sequences 1–50 are marked by "○".

Fig. 2.7. This cluster consists of the sequences labelled by the numbers 10, 11, 12, 13, 14, 20, 21 and 22. This cluster is also visible in Fig. 2.5 (two dark clusters on the diagonal for the sequences 10–14 and 20–22 and two clusters outside the diagonal for the distances between sequences 10–14 and 20–22). The corresponding entries of the distance matrix compose the sub-matrix shown in Table 2.4.

In order to give meaning to these absolute numbers, we calculated the average of the distances between any two random sequences

$$\bar{D} = \frac{1}{N \cdot (N-1)} \sum_{i,j=1}^{N} D(C_i, C_j), \tag{2.35}$$

and, using the random sequences underlying Fig. 2.5 ($N = 50$), we get the numerical value $\bar{D} = 30.9$. Thus, Table 2.4 indicates that all these sequences are indeed similar to each other, which is also hinted by their similar length (not shown). Therefore this set of sequences can be seen as representatives of a *Quasi Species Distribution* (Eigen

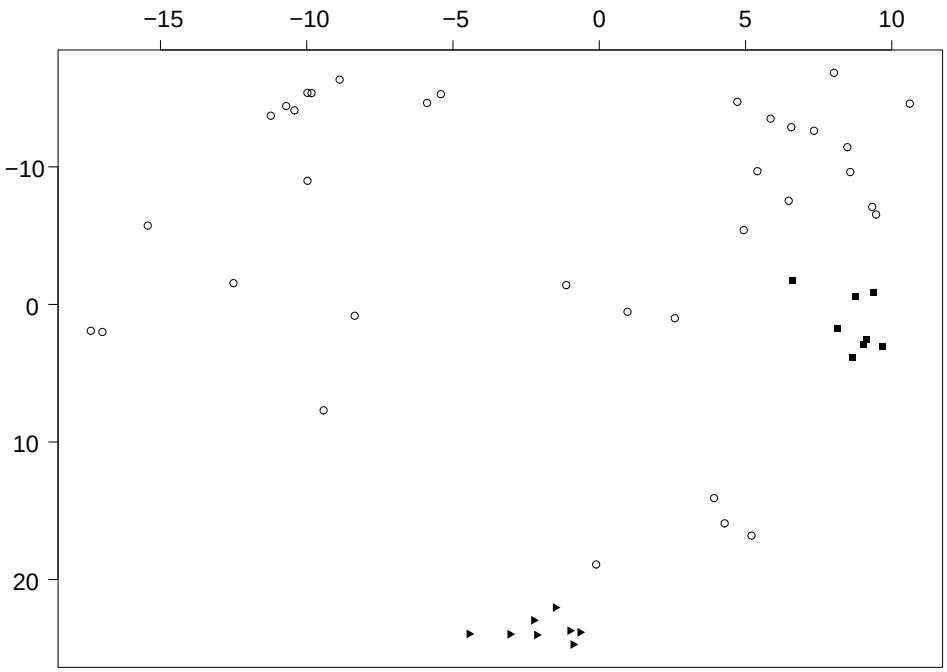


Figure 2.7: Multi-dimensional scaling of the distance matrix for the efficient sequences 1–50. The sequences belonging to cluster 1 (2–9) are marked by a filled square, sequences belonging to cluster 2 are marked by filled triangles.

et al. 1989). In a similar manner, the sequences with numbers 2–9 can be identified as a cluster, representing another *Quasi Species Distribution*. As these sequences yield a higher flux than sequences from the above mentioned cluster, we assume that they are representatives of the *Quasi Master Species Distribution* (see sequences 2–9 in Table 2.2).

In order to find the common properties of sequences inside a cluster, we define a measure of the internal ordering p ($p_{xy} : \mathcal{S} \mapsto [-1, 1]$), with $x, y \in \{\text{A,a,P,p,H,h,N,n}\}$ and $x \neq y$ denoting different generic coupling reactions. The value of this function signifies the *arrangement* of the reactions x and y inside a reaction sequence. The

D	10	11	12	13	14	20	21	22
10	0	4	6	12	8	12	8	10
11	4	0	2	8	4	8	4	6
12	6	2	0	6	2	10	6	4
13	12	8	6	0	4	12	8	6
14	8	4	2	4	0	8	4	2
20	12	8	10	12	8	0	4	6
21	8	4	6	8	4	4	0	2
22	10	6	4	6	2	6	2	0

Table 2.4: Distance sub-matrix of sequences contained in cluster 2. The average distance of two randomly picked sequences is $\bar{D} = 30.9$.

function p_{xy} has the following properties:

$$p_{xy}(C) = 1 \quad \text{if } \textit{all} \text{ reactions of type } x \text{ occur } \textit{earlier} \text{ in the reaction sequence } C \text{ than any reaction of type } y \tag{2.36}$$

$$p_{xy}(C) = 0 \quad \text{if there is no preference which of the reaction types } x \text{ or } y \text{ occurs first} \tag{2.37}$$

$$p_{xy}(C) = -1 \quad \text{if } \textit{all} \text{ reactions of type } x \text{ occur } \textit{later} \text{ in the reaction sequence } C \text{ than any reaction of type } y \tag{2.38}$$

For all intermediary values, $p_{xy}(C)$ reflects the tendency of reactions of type x occurring before reactions of type y ($p_{xy}(C) \geq 0$) or vice versa ($p_{xy}(C) \leq 0$). The exact definition is given in Appendix A.4.2.

Fig. 2.8 shows the p_{xy}-values for all possible combinations of x and y, averaged over three sets of sequences. The black bars denote the corresponding averages for the cluster representing the *Quasi Master Species Distribution* (sequences 2–9, from now on called "cluster 1"), the grey bars for the first mentioned cluster (sequences 10–14, 20–23, from now on called "cluster 2"), the white bars denote the averages for random sequences. Note, that because of the finite lengths of the pathways and due

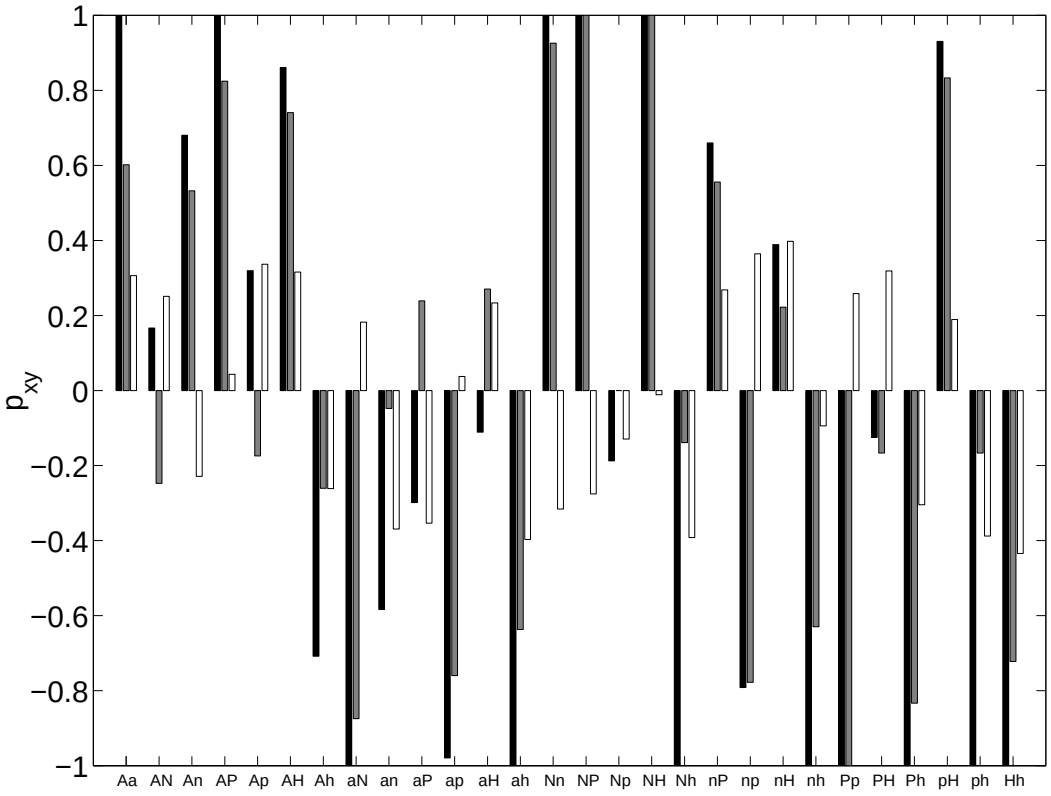


Figure 2.8: Internal ordering of generic reactions. Black bars stand for the *Quasi Master Species Distribution* (cluster 1), grey bars for another quasi species (cluster 2), white bars for random sequences.

to other stoichiometric constraints when generating sequences, the average value for random sequences generally differs from zero. We see, that for some combinations of generic reactions x and y the p_{xy}-values for efficient and random sequences differ greatly, whereas for other combinations the differences are small. From these differences we can conclude which structural features of a pathway are important for its efficiency. Especially we can determine, which reactions are favourable to occur at the beginning of a sequence and which are better located near the end.

Examining particularly ATP consuming (A) and ATP producing (a) reactions, we see that our model reproduces one important feature of glycolysis. All sequences in cluster 1 have a value $p_{Aa} = 1$, meaning that all ATP consuming reactions are located upstream of ATP producing ones. However, sequences of cluster 2, yielding only a marginally smaller flux than those of cluster 1, seem to possess a different internal

structure. The average value $p_{Aa} < 1$ lies between the corresponding values for cluster 1 and random sequences. Note, that in sequence 10 (Table 2.2) which belongs to cluster 2, several 'a'-reactions occur earlier than some 'A'-reactions.

Focusing our interest on NADH consuming (N) and producing (n) reactions, we note that for both clusters 1 and 2, $p_{Nn} \approx 1$, whereas for random sequences this value is negative. This means that NADH producing reactions are located near the end of efficient pathways. This is in agreement with the fact that the citric acid cycle, which produces NADH, occurs later (in the sense of an unbranched reaction chain) than glycolysis, which does not produce NADH.

These results let us formulate the following statement: Of all chemically feasible alternative pathways, those pathways with properties of the internal structure also found in real glycolysis and citric acid cycle are among the most efficient ones with respect to ATP production rate. However, there seem to exist other alternative pathways that are able to produce ATP with almost the same output rate.

2.4 Discussion

Although metabolic pathways are a traditional subject of mathematic modelling, the explanation of the design of these systems is a still unsolved problem. This is partly due to the fact that any understanding of the structural properties requires the consideration of the "evolutionary history" of metabolic systems. One approach is to consider the structure of metabolic systems as the optimal outcome of mutation and selection processes. Whereas much work has already been done concerning the optimisation of the catalytic efficiency of individual enzymes (e. g. Albery and Knowles 1976; Mavrovouniotis and Stepanopoulos 1990; Heinrich and Hoffmann 1991; Pettersson 1992; Wilhelm et al. 1994), studies on the evolutionary optimisation of the stoichiometry of metabolic pathways are still rare.

In the model described in the present chapter, by providing rules for the construction of alternative stoichiometries, we allowed for an optimisation analysis of a vast number of reaction chains. The systematic construction of alternative pathways was made possible by defining classes of reactions (*generic reactions*). Thus it was possible to retrieve stoichiometric properties as the *result* of an optimisation process instead of using this information as a prerequisite as the classical modelling approach does.

The present model is an extension of previous investigations by Heinrich et al. (1997), Meléndez-Hevia et al. (1997), and the model presented by Stephani and Heinrich (1998) concerning the structural design of ATP producing pathways. This ap-

proach can be further extended to analyse the structure of other metabolic systems. For each system an appropriate set of generic reactions has to be defined and, depending on the structure of the intermediates, the mutation rules will have to be adjusted correspondingly. More problematic, however, is the determination of the optimisation criterion. Usually, the "biological purpose" of a certain sub-network of the cellular metabolism is far from being obvious. Therefore, defining a mathematical expression for the performance function is perhaps the most crucial problem when analysing metabolic systems with the presented approach. In principle, it is possible to extend our method to more complex network structures. The set of generic reactions will have to be changed in such a way that they do not only represent certain mono- and bimolecular reactions which may be arranged to unbranched chains but also into branched network structures. In the latter case, several independent steady state fluxes may contribute to the performance of the system which may lead eventually to a multivariate optimisation problem. A systematic approach to studying the structural properties of branched networks is presented in the following chapter.

The main results of our analysis can be summarised as follows. Under the given model assumptions the design of ATP producing pathways characterised by the existence of ATP consuming reactions at the beginning and ATP producing reactions at the end of the pathway is a typical outcome of the optimisation process. Also typical is the occurrence of NADH producing reactions near the end of the pathways. The correspondence between the real aerobic ATP producing system in living cells (glycolysis and TCA cycle) even extends to the number of produced NADH molecules per consumed molecule glucose.

However, within these general features valid for all optimised pathways, variations are still possible. We found several "clusters" of optimal stoichiometries, each of which contains sequences very similar to each other. Comparing sequences belonging to different clusters show significantly greater differences. This means that our results do not exclude the possibility of different designs for ATP producing pathways with almost the same biological efficiency concerning their ATP output rate.

Our results demonstrate that at the derivation of optimal stoichiometries also the kinetic and thermodynamic properties of the individual reactions and of the whole metabolic systems have to be taken into account. This is illustrated, for example, by the fact that in real glycolysis, and also in the optimal sequences listed in Tables 2.2 and 2.3, ATP is invested at the beginning, although the biological function of the pathway as a whole is the production of ATP (see also Heinrich et al. 1997; Teusink et al. 1998). Such an arrangement can only by understood by considering its kinetic advantage compared to other possibilities.

Chapter 3

Branched network structures

In the previous chapter a model has been described analysing the stoichiometric properties of an unbranched reaction chain regarding optimal ATP production. In general, hitherto studies on optimising the kinetic and stoichiometric properties of metabolic reaction systems were restricted mainly to linear arrangements of reactions, i. e. to unbranched pathways. In the present chapter we lift this restriction and allow for networks with branched and cyclic structures. The analysis makes profit from methods developed in the field of metabolic flux analysis in complex networks (Schuster and Hilgetag 1994; Schuster et al. 1999; Schilling and Palsson 1998; Schilling et al. 1999) and makes also use of graph theoretical methods which gain increasing interest in the field of modelling metabolic and regulatory pathways (see Wagner and Fell 2001; Strogatz 2001; Binder and Heinrich 2002).

Inspired by frequently occurring stoichiometric motifs of enzymatic reactions taking place in cellular metabolism our analysis is based on a class of reactions splitting and merging carbon containing compounds. Whereas similar reactions have already been considered before for assembling metabolic pathways (Meléndez-Hevia and Torres 1988; Mittenthal et al. 1998; Mittenthal et al. 2001), we present for the first time a complete analysis with respect to the whole class of possible network structures. Moreover, we give a clear cut definition of "network function" in terms of the capability of a metabolic pathway to perform specific chemical conversions. This allows us to introduce the new concept of "multifunctional networks" which turns out to be of importance for understanding metabolic conversions under varying external conditions. By considering the effect of changes in the composition of networks, similarities in the structures and functions between networks can be analysed in a quantitative way. Describing the resulting changes in terms of mutations enables us to develop the concept

of stoichiometric robustness.

The networks considered in the present paper have been used as a basis to perform evolutionary simulations (see section 1.3) under both fixed and changing environmental conditions. In this way the concept of quasi-species which was developed for describing mutations and selection of macromolecules (Eigen 1971) is further extended to characterise self-organisation of metabolic networks (Heinrich and Sonntag 1981).

3.1 Model assumptions and basic notations

In this chapter the focus lies on the structural design of metabolic networks performing changes in the carbon skeleton of biochemical intermediates as they occur in many pathways in real cells. The compounds C_i participating in the metabolic processes are characterised solely by the number i of their carbon atoms. The set of reactions is confined to processes that alter these numbers. For the sake of keeping the computational effort low, most calculations are performed with a maximal number of carbon atoms $L = 6$ which includes the description of sugars up to hexoses. The generalisation of the model to higher L values allowing the description of longer chained sugars is easily possible.

The L compounds C_i can participate in reversible chemical reactions for which in the following two types are allowed.

- *Reactions of type 1*: Reactions of this type transfer a group of carbon atoms between compounds C_i and C_j resulting in the compounds C_k and C_l. Conservation of the number of carbon atoms in any reaction implies the relation $i + j = k + l$. These reactions $C_i + C_j \rightleftharpoons C_k + C_l$ are bimolecular in both directions, and are called in the following *bi-bi-reactions*. The stoichiometry of every such reaction can be characterised by an integer four-tuple which we write as $(i, j|k, l)$. To represent a chemical reaction the indices must fulfil the condition $(i, j) \neq (k, l)$. Without loss of generality we assume $i \leq j$, $k \leq l$, and $i \leq k$. Physically, any reaction $(i, j|k, l)$ can occur in two ways differing in the number of carbon atoms which are transferred from C_j to C_i. In the first case a group of $l - i = j - k$ carbons and in the second case a smaller group of $k - i = j - l$ carbons is transferred.

- *Reactions of type 2*: Reactions of this type split a compound C_j into two compounds C_k and C_l with $j = k + l$. In the reverse direction they join two compounds. These reactions $C_j \rightleftharpoons C_k + C_l$ are monomolecular in one direction

and bimolecular in the other and are in the following called *bi-uni-reactions*. In order to unify the mathematical description of these two types of reactions it is convenient to formally introduce a compound C_0 having no carbon atom. In this way reactions of the second type can be written as $C_0 + C_j \rightleftharpoons C_k + C_l$ and are characterised by the four-tuple $(0, j|k, l)$.

This model allows for the description of many reactions occurring for example in glycolysis, pentose phosphate pathway, citric acid cycle, and Calvin cycle. A detailed overview of such reactions is given in section 3.1.2.

For $L = 6$ there exist a total of $R = 22$ possible reactions which are listed in Table 3.1. Out of these reactions, 13 are of type 1 and nine of type 2 (see Table 3.1).

reactions of type 1	
number	reaction
1	$(1,3\|2,2)$
2	$(1,4\|2,3)$
3	$(1,5\|2,4)$
4	$(1,5\|3,3)$
5	$(1,6\|2,5)$
6	$(1,6\|3,4)$
7	$(2,4\|3,3)$
8	$(2,5\|3,4)$
9	$(2,6\|3,5)$
10	$(2,6\|4,4)$
11	$(3,5\|4,4)$
12	$(3,6\|4,5)$
13	$(4,6\|5,5)$

reactions of type 2	
number	reaction
14	$(0,2\|1,1)$
15	$(0,3\|1,2)$
16	$(0,4\|1,3)$
17	$(0,4\|2,2)$
18	$(0,5\|1,4)$
19	$(0,5\|2,3)$
20	$(0,6\|1,5)$
21	$(0,6\|2,4)$
22	$(0,6\|3,3)$

Table 3.1: List of all possible reactions with participating reactants with maximal number of carbon atoms $L = 6$. Reactions 1 – 13 are bimolecular in both directions (type 1), reactions 14 – 22 are monomolecular in one direction and bimolecular in the other (type 2). The reactions are characterized by the number of carbon atoms of the participating compounds.

In the following, these reactions will be used as building blocks for alternative networks containing internal and external compounds and a given number r of reactions. The internal compounds are produced as well as consumed by one or more reactions. The concentrations of external substances are considered to be constant, i. e. they are buffered by environmental processes. Under non-equilibrium conditions there exist net fluxes converting external substrates into external products of the network. The analysis is confined to networks with two external substances, denoted in the following by C_a and C_b. Accordingly, the overall stoichiometric balance of the network requires that b molecules of compound C_a are converted into a molecules of compound C_b or vice versa. We characterise any of these overall conversions by the symbol $\langle a, b\rangle$ where without loss of generality $a > b$. Accordingly, there are $L(L-1)/2$ different overall conversions, i. e. 15 for $L = 6$.

The kinetic properties of each network are governed by a differential equation system having the form

$$\frac{dC}{dt} = \mathbf{N} \cdot V \tag{3.1}$$

where C denotes the vector of the concentrations of all L reactants C_i $(i = 1, \ldots, L)$ as well as of the formal reactant C_0. V represents the vector of the rates of the r reactions participating in the network. $\mathbf{N}$ denotes the stoichiometric matrix containing $L + 1$ rows and r columns. The column corresponding to a reaction $(i, j|k, l)$ contains "-1" in the rows i and j and "1" in the rows k and l and zero elsewhere. In case $i = j$ or $k = l$ the corresponding entries are "-2" or "2". In case that a compound C_j is not involved in any reaction of the network all elements of the j-th row of $\mathbf{N}$ are zero.

Under steady state conditions the production and consumption of all internal metabolites C_i with $i \neq a, b$ are balanced leading to constant concentrations of these metabolites. In contrast to that, external metabolites are consumed or produced by the reactions of the network. Accordingly, the steady state of a network converting C_a into C_b is characterised by the equation

$$\mathbf{N} \cdot V = S \tag{3.2}$$

where $S = (s_0, \ldots, s_L)^T$ is a vector with the elements $s_a = -b$, $s_b = a$ (or multiples thereof) and $s_i = 0$ for $i \neq 0, a, b$. The element s_0 is not vanishing when there is formally a net production or net consumption of the compound C_0.

In the following we concentrate on networks which are elementary with respect to a conversion $\langle a, b\rangle$. By the property "elementary" we mean that the steady state condition (3.2) of the corresponding network cannot be met after elimination of any

reaction. If a network can perform a conversion in a non-elementary way, one or more reactions can be eliminated without destroying the ability to perform that conversion. This results in a smaller network which is elementary with respect to this conversion. A network which is elementary with respect to at least one conversion is called an "elementary network". In Appendix B.1 it is proven that an elementary network can maximally contain $r = L - 1$ reactions.

A given elementary network may perform only one special conversion or several different conversions in an elementary way. All these conversions are in the following called "functions" of that network. Networks which are not elementary with respect to any conversion are called non-functional networks. There are two cases of such networks, either they can perform one or more conversions in a non-elementary way, or they cannot perform any conversion.

3.1.1 Visualisation of networks

For the visualisation of networks and their functions in the following a graphical representation will be used which takes into account the distribution of steady state fluxes over the participating reactions. Specifically, the absolute values of the elements of the solution vector V of Eq. 3.2 define how often the corresponding reaction appears in the graphical representation and the signs define the directions (for positive entries in V the corresponding reactions are plotted from left to right). One and the same network may have different graphical representations since solutions V of Eq. 3.2 may exist for different conversions $\langle a, b \rangle$. External metabolites C_a and C_b occur as "free ends" in this representation, meaning that there is a net production or consumption of these compounds. Internal metabolites are balanced under steady state conditions which means that they are produced and consumed with the same rate. However, even internal metabolites may occur as free ends in the representations. In this case, they are present in the same number as sources (compounds which only serve as input for a reaction) as they are present as sinks (compounds which only occur as output of a reaction). This is simply a matter of convenience, since the topography of the networks can become very complex such that simple pictures cannot always be found. Even for one and the same conversion $\langle a, b \rangle$, there are often a large number of possibilities to visualise the network.

An algorithm has been devised that automatically generates several graphical representations of a network and a specified conversion. The heart of the algorithm is a recursive routine that attaches one reaction to an already existing picture. The subrou-

tine tries all possible ways to attach a reaction, accepts or discards the result in a way the programmer may specify, and calls itself with the newly established picture until no reactions are left to attach. The functioning of the algorithm can schematically be described in a diagram depicted in Fig. 3.1. In this figure the complex decision process

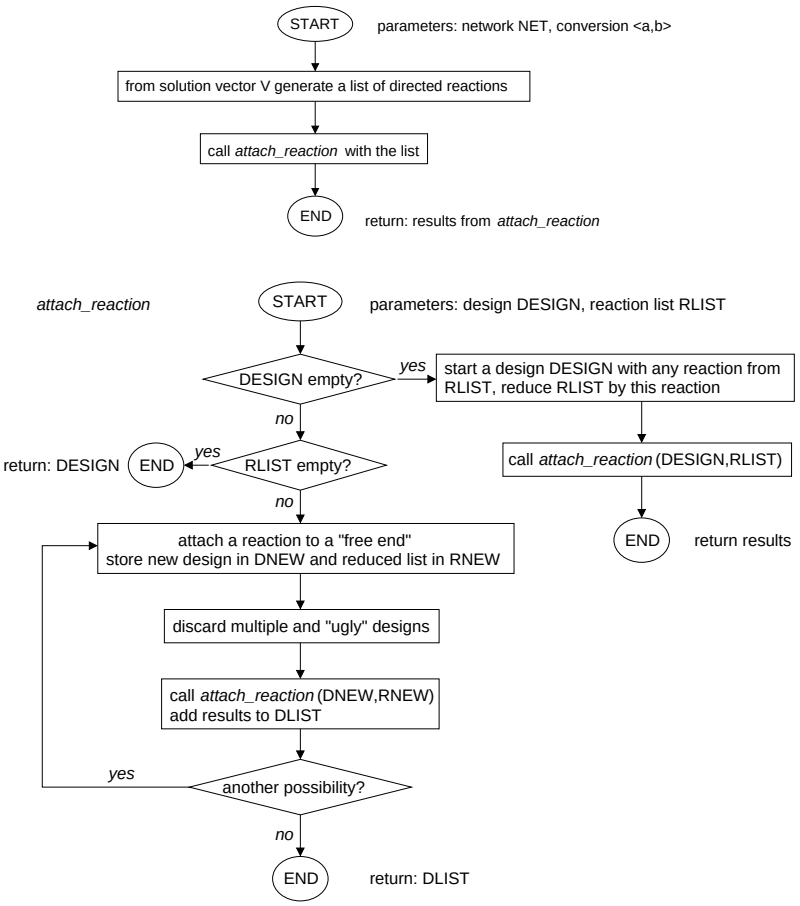


Figure 3.1: Schematic representation of the algorithm to construct graphical representations of a network performing a given function.

which designs shall be discarded and which designs shall be accepted has been comprised in the statement *discard "ugly" designs*. A lot of experimenting was necessary and it can be said that there exists no ideal solution to automate the decision process. However, the algorithm returns a list of possible designs from which the user can choose one to his liking. The explicit placing of the graphical objects representing the reactions and participating compounds and the connecting lines is then performed by another algorithm which essentially minimises the distances between the different ob-

jects. Since the designs are returned as objects, the user has the possibility to perform adjustments by hand a posteriori.

As an example of the outcome of the algorithm, Fig. 3.2 shows a graphical representation of a network of three reactions which can perform the conversion $\langle 6, 1\rangle$ in an elementary way. In this case the only free ends correspond to the external com-

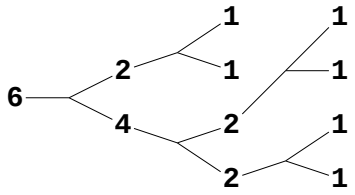


Figure 3.2: Graphical representation of the network consisting of the reactions $(0, 2|1, 1)$, $(0, 4|2, 2)$, and $(0, 6|2, 4)$ with the function $\langle 6, 1\rangle$.

pounds C_6 and C_1. Fig. 3.3 demonstrates the fact that one network performing a function can be displayed in different ways. Here, two graphical representations of the

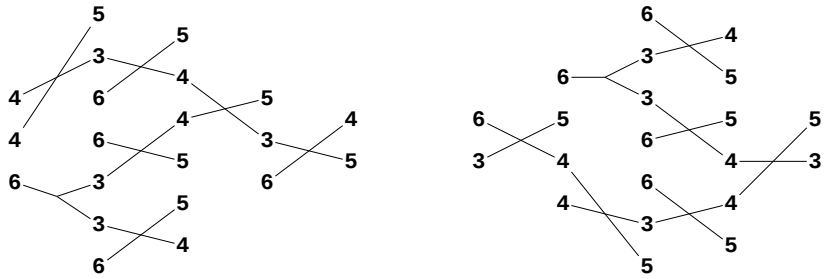


Figure 3.3: Two graphical representations of the network of size $r = 3$ consisting of the reactions $(0, 6|3, 3)$, $(3, 5|4, 4)$, and $(3, 6|4, 5)$ with the function $\langle 6, 5\rangle$.

same network of three reactions which can perform the function $\langle 6, 5\rangle$ in an elementary way are given. In the left representation, the internal compound C_4 appears twice as a source and twice as a sink. In the right representation, the internal compounds C_3 and C_4 both appear once as source and once as sink. In both cases, of course, the external metabolite C_6 appears five times as a source and the external metabolite C_5 appears six times as a sink. This must be read that five six-carbon structures are converted into six five-carbon structures. From now on, at most one representation will be given per function.

3.1.2 Carbon skeleton changing reactions in biological systems

A striking feature of central metabolic pathways in both catabolic and anabolic metabolism is that sugars with different numbers of carbon atoms can be interconverted in a very flexible way. Depending on supply and demand these reactions are regulated according to the needs of the organism.

In order to perform the interconversion of molecules with different numbers of carbon atoms, enzyme catalyzed reactions are used that resemble the reactions defined in this paper as building blocks for metabolic networks.

In glycolysis, the most central part of the energy metabolism in almost every species, the enzyme fructose bisphosphate aldolase splits fructose-1,6-bisphosphate, a hexose (sugar with six carbons) into glyceraldehyde-3-phosphate and its isomer glycerone phosphate, two trioses (sugars with three carbons). In our formalism, this reaction is represented by $(0, 6|3, 3)$. By other enzymatic reactions, glyceraldehyde-3-phosphate is converted into another triose, pyruvate. The pyruvate dehydrogenase complex, a complex containing multiple copies of three enzymes, five coenzymes and two regulatory proteins (see e. g. Alberts et al. 1994), decarboxylases pyruvate to form carbon dioxide (a one-carbon molecule) and acetyl-CoA, which provides the acetyl group (a two-carbon structure) to fuel the citric acid cycle. So in our present formalism this reaction corresponds to $(0, 3|1, 2)$.

Mediated by the enzyme citrate synthase, in the citric acid cycle the acetyl group of acetyl-CoA is joined with oxaloacetate to form citrate (a six-carbon molecule), a reaction corresponding to $(0, 6|2, 4)$. Essentially, this product is decarboxylated twice. First, simultaneously reducing NAD^+ to NADH, isocitrate dehydrogenase catalyses the decarboxylation of isocitrate to form CO_2 and 2-oxo-glutarate. In a second decarboxylation step catalyzed by the 2-oxo-glutarate dehydrogenase system CO_2 and succinyl-CoA is formed. These reactions correspond to $(0, 6|1, 5)$ and $(0, 5|1, 4)$. The description of the central energy metabolism consisting of glycolysis and the citric acid cycle within the presented model is covered in section 3.5.6.

All examples given above represent bi-uni-reactions. However, a number of enzyme catalyzed bi-bi-reactions are present in the pentose phosphate pathway. This pathway provides pentoses (five carbon sugars) and tetroses (four carbon sugars) for biosynthetic processes. These sugars are produced by converting hexoses in a number of steps. Fructose-6-phosphate (a hexose) and glyceraldehyde-3-phosphate (a triose, which is provided by the glycolytic pathway, see above) are transformed

into erythrose-4-phosphate (a tetrose) and xylulose-5-phosphate (a pentose), which corresponds to the bi-bi-reaction $(3, 6|4, 5)$. This process is catalyzed by the enzyme transketolase. This enzyme generally has the property to catalyse the transfer of a two carbon group from one molecule to another. Consequently, other reactions are catalyzed by the same enzyme. In the pentose phosphate pathway, it mediates the conversion of ribose-5-phosphate and xylulose-5-phosphate (two pentoses) into sedoheptulose-7-phosphate (a seven-carbon sugar - a heptose) and the triose glyceraldehyde-3-phosphate, i. e. $(3, 7|5, 5)$. Catalyzed by the enzyme transaldolase, which transfers a three carbon subunit from one compound to another, these two products are converted into fructose-6-phosphate and erythrose-4-phosphate, corresponding to $(3, 7|4, 6)$. However, as the analysis is restricted to a maximal number of $L = 6$ carbon atoms, a representation of these reactions is not included in the model calculations.

In plants, these enzymes also form part of the Calvin cycle, in which CO_2 is fixated for biomass accumulation. The key step which performs the fixation is catalyzed by the enzyme ribulose bisphosphate carboxylase, which lets ribulose-1,5,-bisphosphate react with carbon dioxide to form two molecules 3-phosphoglycerate (a triose). Thus, this reaction resembles $(1, 5|3, 3)$, another bi-bi-reaction.

Summarising it can be concluded that reactions changing the numbers of carbon atoms in the participating compounds play an essential role in cellular metabolism. As well for the energy metabolism (glycolysis and citric acid cycle) as for the production of key compounds used for biosynthesis (pentose phosphate pathway, Calvin cycle in plants), such reactions are indispensable. It is therefore of interest to study the design of metabolic systems making use of these reactions.

3.2 Basic properties and network functions

3.2.1 Number of networks

An algorithm has been developed with which all elementary networks can be constructed and stored into a database. This way, it is easy to determine the according numbers $P_r(a, b)$ of elementary networks for all possible conversions $\langle a, b \rangle$ and network sizes r. The basic principles of the algorithm is depicted in Fig. 3.4. In this algorithm,

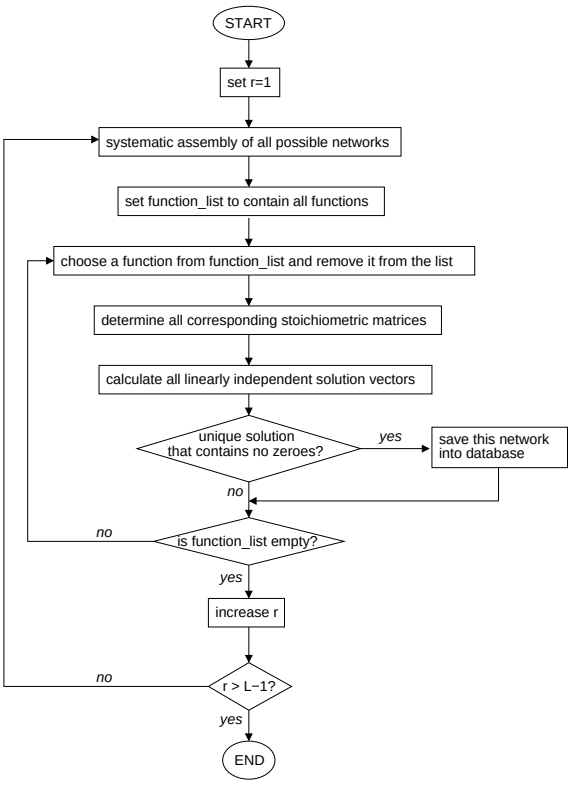


Figure 3.4: The algorithm by which all elementary networks have been identified.

the condition whether a solution V of Eq. (3.2) is unique and does not contain a zero element ensures that the regarded network is elementary, which can be seen as follows:

- if a solution V has a zero element, the corresponding reaction is not needed to perform the given conversion which means that a network is not elementary;
- if Eq. (3.2) has more than one independent solution V they can be linearly combined in such a way that the resulting vector has at least one zero element.

The algorithm has been applied to networks whose participating compounds possess not more than $L = 6$ carbon atoms. The results are presented in Table 3.2. In

network function	network size r				
	1	2	3	4	5
$\langle 6,5\rangle$	–	–	6	379	9146
$\langle 6,4\rangle$	–	3	32	438	4237
$\langle 6,3\rangle$	1	4	35	285	1952
$\langle 6,2\rangle$	–	3	32	438	4237
$\langle 6,1\rangle$	–	–	6	379	9146
$\langle 5,4\rangle$	–	–	20	590	6328
$\langle 5,3\rangle$	–	–	22	580	6247
$\langle 5,2\rangle$	–	–	22	581	6212
$\langle 5,1\rangle$	–	–	14	541	7381
$\langle 4,3\rangle$	–	1	42	526	4497
$\langle 4,2\rangle$	1	4	40	309	1833
$\langle 4,1\rangle$	–	1	31	526	5499
$\langle 3,2\rangle$	–	4	45	465	3082
$\langle 3,1\rangle$	–	3	41	487	3957
$\langle 2,1\rangle$	1	4	42	306	1933

Table 3.2: The total number $P_r(a,b)$ of elementary networks for each function $\langle a,b\rangle$ depending on the network size r.

the first column the trivial case $r = 1$ is presented which includes the three possible conversions $\langle 6,3\rangle$, $\langle 4,2\rangle$ and $\langle 2,1\rangle$ which can be performed with a single reaction. Networks consisting of two reactions cover a wider range of functions: For 9 out of all 15 conversions there are such networks performing this conversion in an elementary way. For $r > 2$ one may always find several networks fulfilling the task to convert C_a into C_b for any given a and b in an elementary way. It is seen that for all functions the numbers $P_r(a,b)$ increase strongly with increasing r.

Closer inspection of Table 3.2 reveals a symmetry, i. e. the numbers of networks for functions $\langle 6,1\rangle$ and $\langle 6,2\rangle$ are exactly the same as for $\langle 6,5\rangle$ and $\langle 6,4\rangle$, respectively.

This symmetry originates from the fact that for the given pairs of conversions a one to one relation can be constructed by renumbering the participating compounds (the proof is presented in appendix B.2).

Interestingly, the numbers of networks with five reactions for the functions $\langle 6,3\rangle$, $\langle 4,2\rangle$ and $\langle 2,1\rangle$ are significantly lower than the number of networks with five reactions for the other functions. This may be understood by the fact that these three conversions can be carried out by a single reaction (see Table 3.2) which means that in each of these cases the corresponding reaction must not occur in a larger network, which otherwise would be non-elementary. A similar, but less pronounced reduction of the number of networks is observed for conversions for which at least two reactions are needed, i. e. $\langle 3,1\rangle$, $\langle 3,2\rangle$, $\langle 4,1\rangle$, $\langle 4,3\rangle$, $\langle 6,2\rangle$, and $\langle 6,4\rangle$ – see Table 3.2. This means that in each of these cases any combination of two reactions representing an elementary network with the given function is not allowed to occur in larger networks.

It is worth mentioning that the present concept of elementary networks is, from a mathematical point of view, closely related to the concept of elementary flux modes developed by Schuster and Hilgetag (1994); see also Heinrich and Schuster (1996). Considering a network which consists of all R reactions and setting the compounds C_a and C_b external, the elementary flux modes of this metabolic system coincide with the networks which are elementary with respect to the conversion $\langle a,b\rangle$. However, we will investigate problems where the elementary networks should be considered as separate systems and not as special flux modes of a complete network of the given class of reactions. This concerns, in particular, the evolutionary development of networks in which completely new designs are explored (see section 3.6).

3.2.2 Multifunctional networks

Inspection of the solutions V of Eq. (3.2) for all possible networks (for $L = 6$) shows that many of them possess not only one but several functions. We define a network's degree f of multifunctionality as the number of its functions. Non-functional networks are characterised by $f = 0$.

For an example of a multifunctional network see Fig. 3.20 on page 75. This network is multifunctional of degree $f = 6$ which is the highest possible degree for a network of size $r = 3$ (see below).

Multifunctional networks which are characterised by $f > 1$ are counted in f different rows of Table 3.2. For given r, the total number Q_r of different networks is generally

less than the sum P_r^{tot} of the numbers of functions over all networks of size r,

$$Q_r \leq \sum_{a>b}^{L} P_r(a,b) = P_r^{\text{tot}}. \tag{3.3}$$

It is interesting to compare these numbers with the total numbers $N_r = \binom{N}{r}$ of possible networks including non-functional networks. The three values Q_r, P_r^{tot} and N_r are represented in Fig. 3.5 as functions of the network size r.

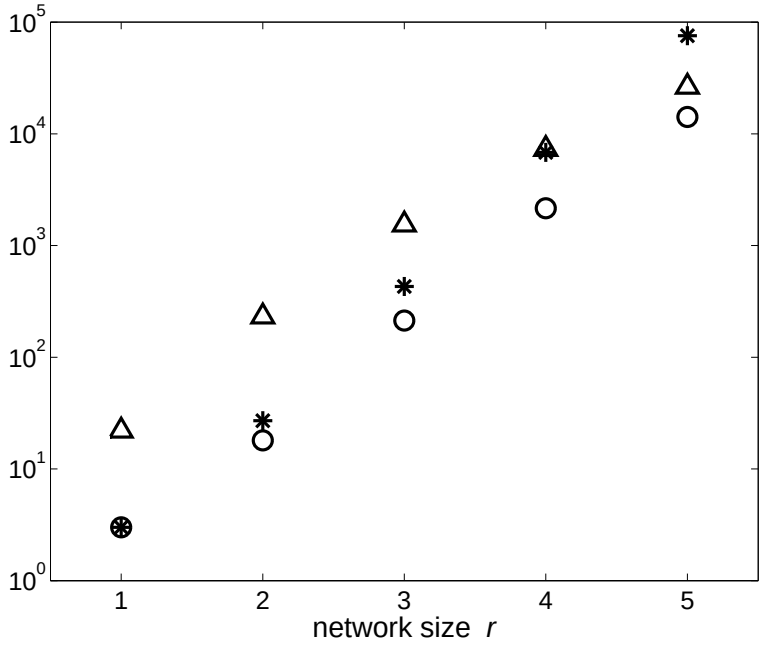


Figure 3.5: The total number Q_r of elementary networks, the sum of the numbers of functions P_r^{tot} over all networks and the total number N_r of all possible networks as functions of r. The values of Q_r are depicted by the symbol $\circ$, the values of P_r^{tot} are represented by the symbol $*$ and the symbol $\triangle$ stands for the values of N_r

An interesting value is the ratio of the first two of these quantities which can be interpreted as the mean degree of multifunctionality:

$$\bar{f}_r = \frac{P_r^{\text{tot}}}{Q_r} \tag{3.4}$$

The upper part of Fig. 3.6 represents the $\bar{f}_r$-values as a function of r. A monotonous increase of the mean degree of multifunctionality with increasing network size r can

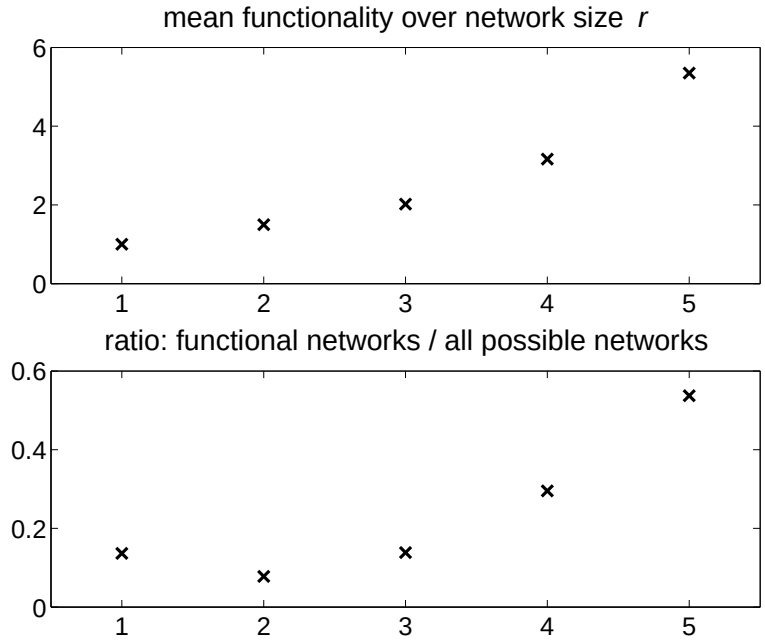


Figure 3.6: The mean degree $\bar{f}_r$ of functionality and the fraction of networks which may perform at least one function depending on the network size r

clearly be observed. Since no single reaction can perform more than one conversion, one obtains $\bar{f}_1 = 1$. Networks of size $r = 5$ show a mean degree of multifunctionality larger than five.

The total numbers Q_r of elementary networks can be compared with the total numbers N_r of networks including non-functional networks. The lower part of Fig. 3.6 represents the ratio Q_r/N_r as a function of network size r. This ratio represents the fraction of all possible networks which may perform at least one function in an elementary way. There is the general tendency that this fraction increases with increasing r. There is, however, the exception that for networks consisting of two reactions this ratio is smaller than for single reaction networks.

The exact values for the quantities Q_r, P_r^{tot}, N_r, $\bar{f}_r$ and Q_r/N_r which were used to generate Figs. 3.5 and 3.6 are presented in Table 3.3.

Table 3.4 lists the numbers of networks of varying size which are multifunctional to a certain degree. It is clearly seen that the ability to act as a multifunctional network increases with increasing size. For example, networks with two reactions possess maximally three functions whereas networks of 5 reactions may possess up to 14 func-

quantity	network size r				
	1	2	3	4	5
Q_r	3	18	213	2160	14152
P_r^{tot}	3	27	430	6830	75687
N_r	22	231	1540	7315	26334
$\bar{f}_r$	1	1.50	2.02	3.16	5.35
Q_r/N_r	0.136	0.078	0.138	0.295	0.537

Table 3.3: Some statistical properties of elementary networks: The total number Q_r of different elementary networks for a given r, the sum P_r^{tot} of the functions over all networks of size r, the total number N_r of networks of size r including non-functional networks, the mean degree $\bar{f}_r = P_r^{\text{tot}}/Q_r$ of multifunctionality for a given r and the ratio Q_r/N_r of elementary networks.

tions. Networks which are elementary with respect to all 15 conversions do not exist. For any given network size r the number of monofunctional networks is always high and the number of networks with maximal degree is always low. There is however, no monotonous relation between the number of networks and the degree of multifunctionality f.

Considering the biological function of a metabolic system such as the interconversion of two metabolic compounds, also those conversions should be taken into account which the network can perform in a non-elementary way. Especially networks which are able to perform more than one conversion should have no disadvantages if they can perform some functions in a non-elementary way. It is therefore of interest to investigate the total numbers of conversions (elementary and non-elementary) which a network can perform that is elementary with respect to *at least one* conversion. These numbers are listed in Table 3.5.

A striking feature of Table 3.5 is that there are gaps in the total numbers of conversions an elementary network can perform. For example, networks consisting of two reactions either can perform one or three conversions, but there exists no network of size two which can perform exactly two conversions. Similarly, there are no networks of size three which can perform exactly four or five conversions: either they can perform less than four or exactly six conversions. For $r = 4$, there are no networks with five, seven, eight or nine conversions. Finally, for networks of size $r = 5$, there even

size r	degree f of multifunctionality													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	3													
2	11	5	2											
3	109	42	37	4	16	5								
4	718	367	301	330	67	119	59	99	99	1				
5	1909	1491	1315	2278	2433	324	312	971	610	807	752	550	354	46

Table 3.4: Numbers of networks having a certain degree f of multifunctionality depending on the network size r.

size r	total numbers of conversions (including non-elementary)														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	3														
2	10		8												
3	76	25	42			70									
4	173	317	383	7		165				1115					
5														14152	

Table 3.5: Numbers of networks that are able to perform a certain number of conversions and are elementary with respect to at least one of these, depending on the network size r.

exist *only* networks that can perform all 15 conversions. In other words, a network of size $r = 5$ that is elementary with respect to *any* conversion is able to perform *all* conversions (in a possibly non-elementary way). This is an astonishing result which is very unlikely to be a coincidence. Indeed, further analysis reveals the following general rule:

> Let L denote the maximal number of carbon atoms within a compound. Any network that is elementary with respect to an arbitrary conversion and which has maximal size (i. e. a network consisting of $L - 1$ reactions – see Appendix B.1) can perform all conversions $\langle a, b \rangle$, $1 \leq a < b \leq L$.

The mathematical proof of this rule is presented in Appendix B.3.

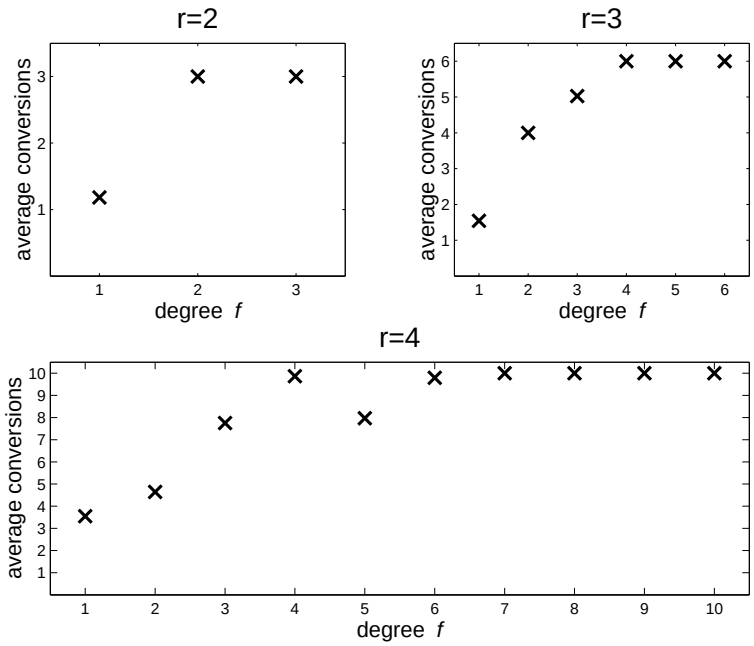


Figure 3.7: The dependency of the average number of conversions on the degree f of multifunctionality.

For smaller network sizes ($r < L - 1$) it is interesting to investigate the relation between the degrees f of multifunctionality and the total numbers of conversions networks can perform. Again assuming $L = 6$, Fig. 3.7 shows the numbers of conversions averaged over all networks of a given size r and a given degree f of multifunctionality. As can be expected, in general an increase of the number of conversions is observed with increasing f. However, the increase is not monotonous. Interestingly, networks of size $r = 4$ with a degree of multifunctionality $f = 5$ can in average perform less functions than networks of the same size but with $f = 4$.

3.3 Composition of the networks

3.3.1 Frequency of the specific reactions

Having identified all elementary networks for $L = 6$, it is of interest to analyse their composition. For each reaction $\mathcal{R}$ and each network size r, the frequencies $n_r(\mathcal{R})$ of the

occurrence of all $R = 22$ reactions within networks of this size have been calculated. For single reaction networks ($r = 1$) the result is trivial. The interesting fact can be observed that not all reactions are found in the set of all networks of size $r = 2$. The absolute numbers of the occurrence of the single reactions are presented in Table 3.6. The reactions $(0, 6|3, 3)$, $(3, 5|4, 4)$, $(3, 6|4, 5)$ and $(3, 6|4, 5)$ cannot be combined with

frequency	reactions
1	$(0, 5\|1, 4), (1, 5\|2, 4), (1, 5\|3, 3), (2, 5\|3, 4), (2, 6\|3, 5), (1, 6\|3, 4), (1, 6\|2, 5)$
2	$(0, 5\|2, 3), (0, 6\|1, 5), (1, 4\|2, 3), (2, 4\|3, 3), (2, 6\|4, 4)$
3	$(0, 2\|1, 1), (0, 3\|1, 2), (0, 4\|1, 3), (0, 6\|2, 4), (1, 3\|2, 2)$
4	$(0, 4\|2, 2)$

Table 3.6: Absolute numbers of the occurrence of the specific reactions in networks of size $r = 2$. Reactions which are not mentioned do not occur.

another reaction to form a functional network.

In order to compare the frequencies $n_r(\mathcal{R})$ for different network sizes, the scaled values $(R/rQ_r) \cdot n_r(\mathcal{R})$ are considered. The scaling factor has been chosen such that

$$\frac{1}{R} \sum_{\mathcal{R}} \frac{R}{rQ_r} n_r(\mathcal{R}) = 1, \tag{3.5}$$

meaning that for each network size r the rescaled values have an average of one. The result has been plotted in Fig. 3.8 where the reactions have been ordered by their total number of occurrence within all elementary networks beginning with the most frequent reaction on the left. It can be clearly seen that the distribution becomes more levelled with increasing network size. However, for a random distribution of frequencies, this is a behaviour that is expected, since the numbers Q_r of elementary networks strongly increase with increasing size r (see Fig. 3.5).

In order to be able to decide how significant the deviations from the mean values are, a χ^2-test has been performed (see e. g. Sachs 1992). This test checks the values against a null hypothesis. Here, the question of interest is whether these distributions of frequencies differ significantly from a distribution that would result from assigning reactions to the networks in a completely random fashion. In other words, the question can be formulated as follows: "can these frequencies be distinguished from the outcome of an experiment in which an unbalanced 22-sided die is thrown rQ_r times?". Therefore, under assumption of the null hypothesis, the expected value $E(\mathcal{R})$ for the frequency is

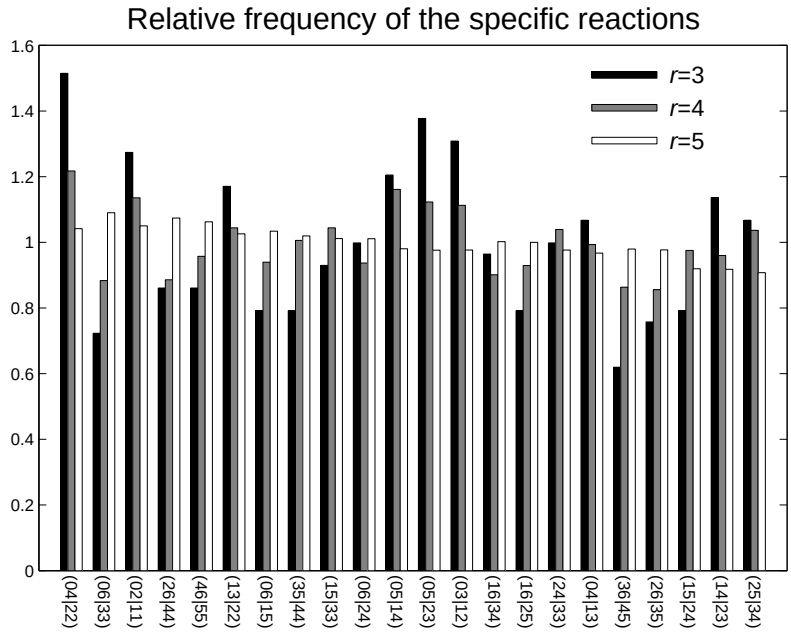


Figure 3.8: The relative frequencies of the reactions as they occur in networks of size $r = 3, 4, 5$.

the same for each reaction $\mathcal{R}$ and amounts to

$$E(\mathcal{R}) = \frac{rQ_r}{R}, \tag{3.6}$$

where r is the size of the networks, Q_r the number of elementary networks with size r and R the total number of different reactions. For each r, the test-value $\hat{\chi}^2$ hat to be calculated using the formula (Sachs 1992)

$$\hat{\chi}^2 = \sum_{\mathcal{R}} \frac{(n_r(\mathcal{R}) - E(\mathcal{R}))^2}{E(\mathcal{R})}. \tag{3.7}$$

The corresponding values are given in Table 3.7.

network size r	3	4	5
$\hat{\chi}^2$	34.117	85.837	161.058

Table 3.7: Test values $\hat{\chi}^2$ for the distributions of the frequencies of the specific reactions within networks of sizes $r = 3, 4, 5$.

These values have to be compared with the known χ^2-distribution for $R - 1 = 21$ degrees of freedom. For the confidence levels of 5%, 0.5% and 0.05% the values are

$$\chi^2(21, 0.05) = 32.671, \tag{3.8}$$

$$\chi^2(21, 0.005) = 41.401, \tag{3.9}$$

$$\text{and} \quad \chi^2(21, 0.0005) = 49.011. \tag{3.10}$$

At least for network sizes $r \geq 4$ it can be said with almost absolute certainty that the distribution of reaction frequencies is not random.

Interestingly, the most abundant reactions are the reactions $(0, 2|1, 1)$, $(0, 4|2, 2)$ and $(0, 6|3, 3)$ which by itself can perform a function, i. e. which form an elementary network of size $r = 1$. The general tendency can be observed from Fig. 3.8 that those reactions are represented stronger which are either bi-uni-reactions or which metabolise two compounds of the same kind – such as the reaction $(2, 6|4, 4)$.

Another observation made from Fig. 3.8 is that in general the relative frequencies of one and the same reaction either increase or decrease with increasing network size. Only for a few reactions the relative frequency for networks of size $r = 4$ is maximal – reactions $(1, 5|3, 3)$ and $(2, 4|3, 3)$ – or minimal – reactions $(0, 6|2, 4)$ and $(1, 6|3, 4)$.

3.3.2 Abundance of pairs of reactions

Another interesting question is how reactions correlate within the networks. In other words: "Are there pairs of reactions which occur significantly often together within a network?". If the answer to this question is positive, one might conclude that there are typical structures consisting of several reactions which are favourable for a network to be able to perform functions.

In order to test for statistical significance, the following procedure has been applied: For each given network size r and each pair of reactions, the numbers of networks have been determined which contain this combination of reactions. These numbers fill an upper triangular matrix.

The null hypothesis claims that there is no correlation between reactions within networks, therefore the numbers in the above constructed matrix should not be distinguishable from completely random numbers (with a given total).

To test against the hypothesis, a Monte-Carlo simulation has been performed. Of course, a similar approach as in the previous section using a χ^2-test is also possible, but the results of the Monte-Carlo simulation prove more intuitive.

As for $R = 22$ reactions the vast majority of networks are of size $r = 5$, the simulation has been carried out for these values as follows: According to the null hypothesis, 1000 data sets have been generated by randomly choosing Q_r sets of r numbers out of the numbers $1, \ldots, R$ representing the R reactions. For each data set, the frequencies of occurring network pairs have been determined resulting in a set of 1000 numbers for each frequency. These 1000 numbers determine an average value and a 5% exclusion range This exclusion range has been determined by identifying the 50 most extreme values out of the 1000 numbers for each frequency. The average (thick curve) and the exclusion limits (thin curve) have been plotted over all occurring frequencies into Fig. 3.9. To this plot the observed values resulting from all elementary networks of

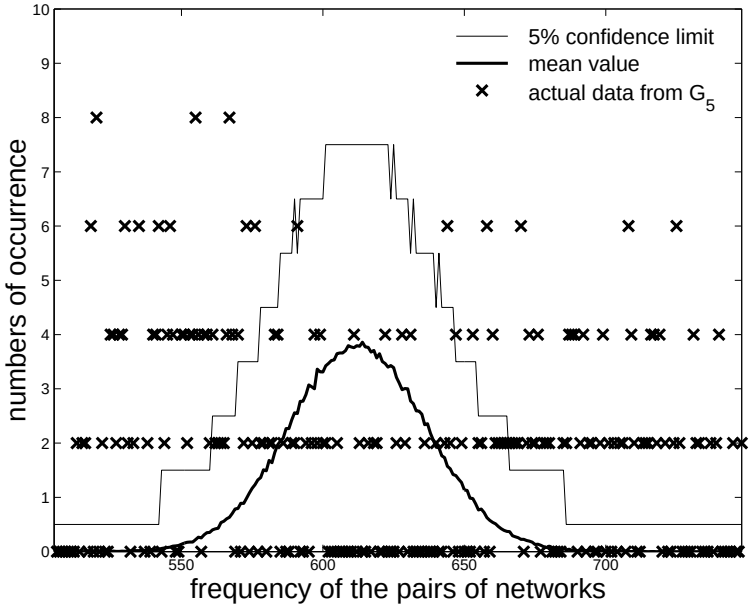


Figure 3.9: Results from a Monte-Carlo simulation to test against the null hypothesis that the reactions occur completely uncorrelated within the networks.

the given size are added (crosses). It is now possible to observe from the plot, how often an observed value lies in the exclusion range and from this observation the null hypothesis can be either rejected or accepted. In the present case, the distribution of the observed values deviates strongly from the random distribution which has been generated according to the null hypothesis. Many values (close to 40%) clearly fall out of the confidence range. Therefore it can be concluded that it is extremely unlikely for the observed values to be in accordance with the null hypothesis. Another interesting

fact seen from Fig. 3.9 is that all frequencies of pairs of networks occur an even number of times. This behaviour, however, has not been further investigated. Nevertheless this is a behaviour absolutely contradicting the hypothesis that the frequencies originate from a random distribution. These observations strongly support the assumption that there is a correlation of reactions.

Similar results have been produced from simulations for network sizes $r = 3$ and $r = 4$. However, the smaller the network size, the less pronounced is the deviation from the null hypothesis.

A closer inspection shows that the most abundant combinations of reactions in networks of size $r = 3$ and $r = 4$ are $(0, 5|1, 4)$ with $(1, 5|3, 3)$ as well as the combination $(0, 4|2, 2)$ with $(2, 4|3, 3)$. Both of these pairs occur ten times in networks of size $r = 3$ and 98 times in networks of size $r = 4$. Networks of size $r = 5$ contain most frequently the combination of the reactions $(0, 6|3, 3)$ and $(1, 3|2, 2)$, which occurs 748 times. It is noticeable that all these examples represent a combination of a bi-uni- with a bi-bi-reaction.

3.4 Network-network relations

3.4.1 Transitions, mutations and distances between networks

Each elementary network consists of a subset of r reactions out of all R possible reactions. We are interested now in the effect of exchanging exactly one reaction for another one not present before in the given network. Such an exchange may result in a transition from an elementary network to another elementary network or to a non-functional network of the same size. Transitions resulting in other elementary networks are in the following called *mutations*. There are different types of mutations which may be classified as follows:

PL loss of some functions without acquiring new functions (partial loss of function),

G acquisition of new functions by maintaining all old functions (gain of function),

CC loss of all previous functions and simultaneous acquisition of new functions (complete change of function),

PC loss of some functions and simultaneous acquisition of new functions (partial change of function), and

N neither loss nor gain of functions (neutral mutation).

Note that if a PL-mutation from one network to another exists then the reverse is a G-mutation. For CC-, PC-, and N-mutations the mutation and its reverse are of the same class.

If an exchange of one reaction results in a non-functional network we call this exchange a CL-transition (complete loss of function).

This concept of mutations allows to define a neighbourhood relation between networks. We call two networks $\mathcal{N}$ and $\mathcal{N}'$ neighbours when they can be interconverted by a single mutation. These neighbourhood relations define a graph G_r for each r in which the networks are the vertices and the possible mutations are the edges. If a network $\mathcal{N}'$ can result from a network $\mathcal{N}$ by a sequence of mutations, these mutations form a path in the graph. The *distance* $d(\mathcal{N}, \mathcal{N}')$ between such two networks is defined as the length of the shortest path connecting the two corresponding vertices. There is the possibility that the graph G_r is not connected, i. e. there exist networks $\mathcal{N}'$ which cannot be reached from a given network $\mathcal{N}$. A distance between such pairs of networks is not defined.

For a connected graph a mean distance between pairs of networks may be defined as follows

$$\bar{d} = \frac{1}{Q_r(Q_r - 1)} \sum_{i,j=1}^{Q_r} d(\mathcal{N}_i, \mathcal{N}_j), \tag{3.11}$$

where the networks are numbered by the indices i and j.

The total number of transitions from one elementary network of size r resulting in other networks amounts to

$$t_r^{\mathrm{tot}} = r(R - r) \tag{3.12}$$

since every one of the r reactions can be replaced in $R - r$ ways. All transitions which are mutations contribute to the connectivities of the vertices defined as the number of their neighbours.

In the following we will consider not only transitions from individual networks but the entirety of the transitions originating from any vertex in G_r. The total number of such transitions amounts to

$$T_r^{\mathrm{tot}} = t_r^{\mathrm{tot}} Q_r = r(R - r) Q_r. \tag{3.13}$$

T_r^{tot} can be written as the sum of numbers of the transitions in the various classes

$$T_r^{\mathrm{tot}} = T_r^{\mathrm{CC}} + T_r^{\mathrm{PL}} + T_r^{\mathrm{G}} + T_r^{\mathrm{PC}} + T_r^{\mathrm{N}} + T_r^{\mathrm{CL}}, \tag{3.14}$$

where $T_r^{\mathrm{PL}} = T_r^{\mathrm{G}}$.

3.4.2 Properties of the graphs G_r

For $L = 6$, the graphs G_r, $r = 1, \ldots 5$ have been generated. Fig. 3.10 shows a representation of the Graph G_2, i. e. the graph that contains all networks of size $r = 2$ as vertices. The 26 edges represent possible mutations between the networks: Two networks which are connected by an edge in G_2 differ in exactly one reaction. Most of the mutations (20) that are possible are CC-mutations. The edges corresponding to the remaining six mutations are labelled by the type of mutations they represent.

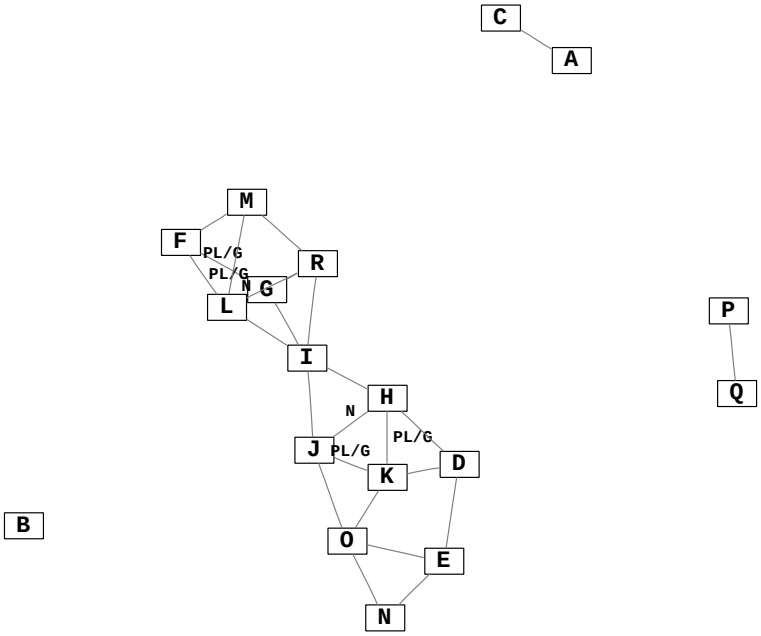


Figure 3.10: The graph G_2.

The graph G_2 consists of $Q_2 = 18$ vertices corresponding to 11 monofunctional, five bifunctional and two trifunctional networks. The vertices are labelled by the letters "**A**" through "**R**" and the composition of all these networks and their functions are given in Table 3.8. The graph G_2 decomposes into 4 components with 13, 2, 2 and 1 networks. This means in particular that two networks represented by vertices belonging to different components cannot be transformed into each other by a sequence of mutations. Consequently, a distance is only defined between such pairs of networks which belong to the same component.

In contrast to G_2, the graph G_3 is connected, i. e. any network can be reached from any other network by a sequence of mutations represented by a path along the

descriptor from Fig. 3.10	composition	functions
A	$(0,5\|2,3)$, $(2,5\|3,4)$	$\langle 4,2\rangle$
B	$(0,5\|1,4)$, $(1,5\|2,4)$	$\langle 2,1\rangle$
C	$(0,5\|2,3)$, $(2,6\|3,5)$	$\langle 6,3\rangle$
D	$(0,3\|1,2)$, $(1,4\|2,3)$	$\langle 4,2\rangle$
E	$(0,4\|1,3)$, $(1,4\|2,3)$	$\langle 2,1\rangle$
F	$(0,6\|2,4)$, $(2,6\|4,4)$	$\langle 4,2\rangle$, $\langle 6,2\rangle$, $\langle 6,4\rangle$
G	$(0,4\|2,2)$, $(2,6\|4,4)$	$\langle 6,2\rangle$, $\langle 6,4\rangle$
H	$(0,2\|1,1)$, $(0,3\|1,2)$	$\langle 3,1\rangle$, $\langle 3,2\rangle$
I	$(0,2\|1,1)$, $(0,4\|2,2)$	$\langle 4,1\rangle$
J	$(0,2\|1,1)$, $(1,3\|2,2)$	$\langle 3,1\rangle$, $\langle 3,2\rangle$
K	$(0,3\|1,2)$, $(1,3\|2,2)$	$\langle 2,1\rangle$, $\langle 3,1\rangle$, $\langle 3,2\rangle$
L	$(0,4\|2,2)$, $(0,6\|2,4)$	$\langle 6,2\rangle$, $\langle 6,4\rangle$
M	$(0,6\|2,4)$, $(2,4\|3,3)$	$\langle 6,3\rangle$
N	$(0,4\|1,3)$, $(1,6\|3,4)$	$\langle 6,3\rangle$
O	$(0,4\|1,3)$, $(1,3\|2,2)$	$\langle 4,2\rangle$
P	$(0,6\|1,5)$, $(1,6\|2,5)$	$\langle 2,1\rangle$
Q	$(0,6\|1,5)$, $(1,5\|3,3)$	$\langle 6,3\rangle$
R	$(0,4\|2,2)$, $(2,4\|3,3)$	$\langle 3,2\rangle$, $\langle 4,3\rangle$

Table 3.8: Composition and functions of all elementary networks of size $r = 2$.

edges (computational analysis reveals that the same holds true for the graphs G_4 and G_5, which are not shown due to their large size). A representation of the graph G_3 is depicted in Fig. 3.11. The graph G_3 contains a total of $Q_3 = 213$ vertices.

This graph G_3 alone – and even less so the larger graphs G_4 and G_5 – does not reveal very much about the interrelation between the networks but rather has to be analysed by determining some characteristic properties.

Fig. 3.12 shows the distribution of the connectivities of the graphs G_3, G_4 and G_5. For G_3, the minimal connectivity amounts to two and the maximal connectivity to 22. Interestingly, the number of networks having only a few neighbours as well as the number of networks with many neighbours is small, a characteristics which

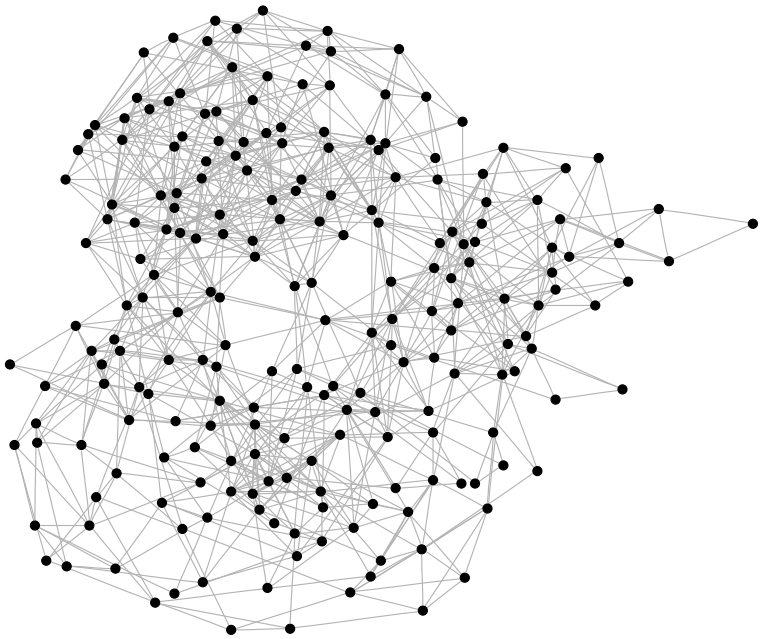


Figure 3.11: The graph G_3.

is also valid for the larger graphs G_4 and G_5. With increasing r, the property that there are only few networks with a small number of neighbours becomes drastically pronounced. For example, the least number of neighbours for networks of size $r = 4$ amounts to six – there are three such networks. Networks consisting of $r = 5$ reactions even have a minimal number of neighbours of 38 (one such network exists). In G_3, the most frequently occurring connectivities are 7, 8, and 9, in G_4, the most common number of neighbours is 23 and in G_5, the most frequent connectivities range from 49 to 51. In general, it can be observed that the shapes of the distribution functions are similar for the three r-values. However, the incline from low connectivities towards the peak of the most frequently occurring connectivities becomes sharper with increasing r, whereas the decline from the peak towards large connectivities becomes more stretched out as the network size r increases. For $r = 5$, there exists a network that has a total of 85 neighbours. The total number t_r^{tot} of transitions (3.12) which is also the theoretical maximum of mutations from one network, evaluated for $R = 22$ and $r = 5$ also amounts to 85. This is an exceptional result meaning that there exists exactly one network which is completely robust towards mutations in the sense that every possible mutation applied to this network results in another functional network. A more detailed investigation of this kind of robustness will be presented in section 3.4.3. Certainly,

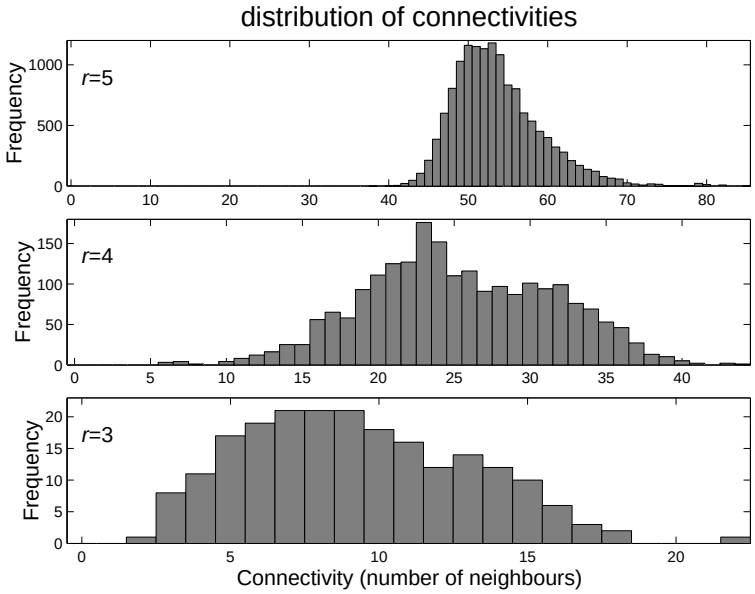


Figure 3.12: Distribution of connectivities in the graphs G_r, $r = 3, 4, 5$.

Fig. 3.12 reveals that the graphs G_r, $r = 3, 4, 5$ are not scale free in the sense proposed by Barabasi and Albert (1999) since the distribution function does not decay as a power law. This is, however, not surprising since the graphs G_r are not the outcome of a process where new edges are preferentially attached to already highly connected nodes. In contrast to investigations on metabolic networks by Jeong et al. (2000), who have shown that the graphs of such networks themselves can be considered to be scale-free, the graphs defined in the present context reflect exclusively the stoichiometric similarities of different network designs. Table 3.9 contains the numbers of edges for each mutation class for networks of sizes $r \geq 2$. The upper part of Fig. 3.13 shows the fraction of the specific mutation types as functions of the network size r. For small networks (sizes $r \leq 4$) the overwhelming number of mutations imply a complete change of function. Considerably less mutations are of the neutral type and partial loss/gain of function mutations. By far the rarest class of mutations for sizes $r \leq 4$ imply a partial change of functions. The high number of complete change of function mutations may be explained by the fact that the majority of networks are of degree $f = 1$ (see Table 3.4) and thus it is very likely for a network to lose its only function as a result of a mutation. For large networks (size $r = 5$) this statement is not valid anymore, there are relatively more networks with degree $f > 1$ and indeed the fraction

mutation class m	number of edges in			
	G_2	G_3	G_4	G_5
CC	20	660	12662	29254
PL/G	4	134	7138	184201
PC	0	30	2720	103009
N	2	165	4800	63373

Table 3.9: Numbers of edges of the graphs G_r, $r = 2, \ldots, 5$ corresponding to the various mutation classes. Note that since every edge represents two mutations (both directions), the corresponding values T_r^m are twice the values given in the table for mutation types $m = \mathrm{CC}, \mathrm{PC}, \mathrm{N}$ and exactly the value for mutation types $m = \mathrm{PL}, \mathrm{G}$.

of CC-mutations drops dramatically to become the least common mutation type. Since most networks of size $r = 5$ possess several functions, they are more likely to keep some of their previous functions after a mutation. Thus, the high proportion of PL/G- as well as PC-mutations can be explained.

It is interesting to compare the sum of all mutations to the number of all possible transitions including CL-transitions. The corresponding ratio is presented as a function of network size r in the lower part of Fig. 3.13. A monotonous increase of the ratio with increasing r is observed. For small networks, the number of CL-transitions is significantly larger than the sum of the numbers of all other transitions. For large networks, there are even more mutations than CL-transitions. This means, the larger a network, the less likely it is for a mutation to cause "severe damage" in the sense that the network loses all of its functions – see section 3.4.3.

For further characterisation of the graphs we consider the distances between the networks. Since G_2 is not connected, a distance is not defined for every pair of networks, but only for networks which belong to the same component. Since G_3 is connected, distances are defined for every pair of networks. Fig. 3.14 represents the distance distribution for the networks of G_3. The distribution is non-monotonous, indicating a low number of directly neighboured networks, a rather higher number of pairs of networks with distances three to four, and again few networks with high distances. The maximal distance between two networks define the diameter of the graph which for G_3 amounts to

$$D(G_3) = 8, \tag{3.15}$$

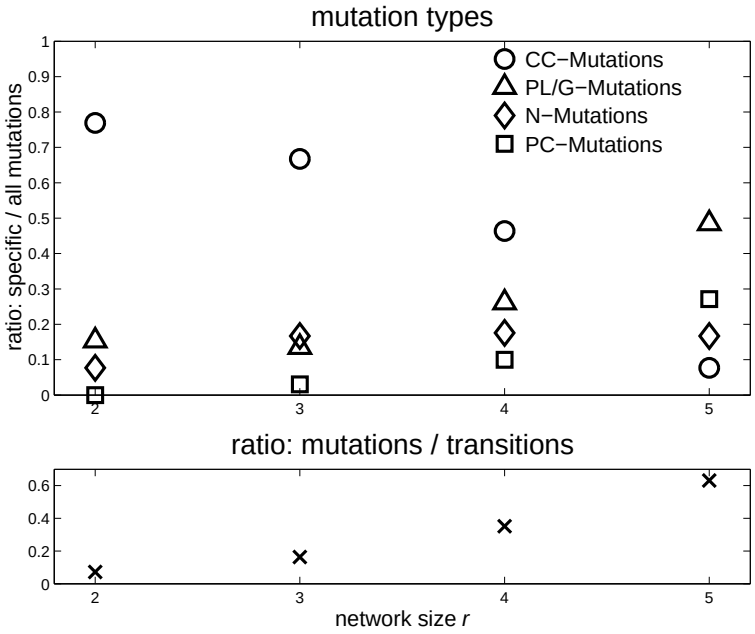


Figure 3.13: Composition of all mutations by mutation types (upper part) and the ratio of possible mutations to all possible transitions including CL-transitions (lower part).

meaning that any two networks can be connected by eight or less mutations. The mean distance between all pairs of networks is $\bar{d}(G_3) = 3.55$.

Mean distances $\bar{d}_{ab}$ can also be defined for subclasses of networks with a given function $\langle a, b\rangle$. The values have been determined and are presented in Table 3.10. Inspection of this Table reveals that most of these values are smaller than $\bar{d}$ (eleven

function	$\langle 6,5\rangle$	$\langle 6,4\rangle$	$\langle 6,3\rangle$	$\langle 6,2\rangle$	$\langle 6,1\rangle$	$\langle 5,4\rangle$	$\langle 5,3\rangle$	$\langle 5,2\rangle$	$\langle 5,1\rangle$	$\langle 4,3\rangle$	$\langle 4,2\rangle$	$\langle 4,1\rangle$	$\langle 3,2\rangle$	$\langle 3,1\rangle$	$\langle 2,1\rangle$
$\bar{d}_{ab}$	2.40	3.83	3.70	3.44	2.00	2.94	2.94	2.59	2.30	3.07	3.61	2.50	3.50	3.38	3.71

Table 3.10: Mean distances of subgraphs $G_3(a, b)$ consisting of networks with a given function.

out of 15). This indicates a tendency towards clustering of networks with the same function within G_3. Since all $\bar{d}_{ab}$-values do not differ very strongly from $\bar{d}$, this clustering is not complete such that networks of the same function are not necessarily close to each other.

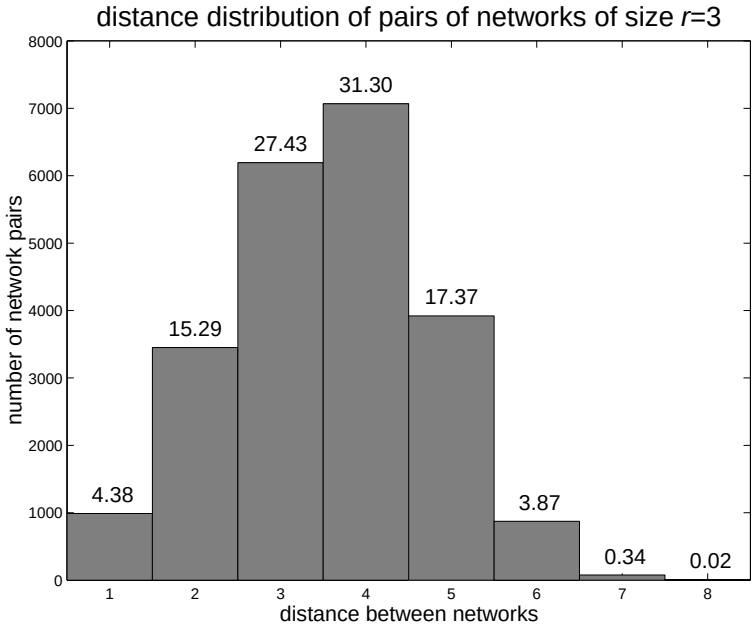


Figure 3.14: The distribution of distances between pairs of networks of size $r = 3$. Above the bars the percentages are given.

These calculations are very time consuming for the graphs of the larger networks of sizes $r = 4, 5$ since the computational effort rises quadratically with the numbers of networks present in the graph. However, an estimation of the diameters $D(G_4)$ and $D(G_5)$ for the graphs G_4 and G_5, respectively, can be obtained as follows: If an arbitrary vertex V is chosen from a graph G, the maximal distance $d_V^{\max}$ from this vertex to another vertex in G is easily obtained, since the computational effort for this kind of calculation only rises linearly with the numbers of vertices. Then the diameter $D(G)$ lies in the range

$$d_V^{\max} \leq D(G) \leq 2d_V^{\max}. \tag{3.16}$$

Obviously, since two networks with distance $d_V^{\max}$ have been identified, the diameter cannot be lower, and, since every two network can be connected by a path of a length not longer than $2d_V^{\max}$ that runs through the vertex V, the upper bound is also immediately clear.

About 100 arbitrary networks have been chosen from both G_4 and G_5 and the $d_V^{\max}$-values have been determined. For G_4, most $d_V^{\max}$-values amount to five, but one such value of four has also been found. Therefore, the diameter of G_4 can be estimated to

lie in the range

$$5 \leq D(G_4) \leq 8. \tag{3.17}$$

For the graph G_5, only vertices resulting in a $d_V^{\max}$-values of five have been identified, therefore the diameter can only be estimated to be

$$5 \leq D(G_5) \leq 10. \tag{3.18}$$

It is remarkable that the diameters of the graphs G_r do not change considerably with increasing network size r. It can be concluded that the connectivities of the graphs increase in such a manner that the average distance stays more or less the same – as far as can be said by the estimations given in Eqs. (3.17) and (3.18) – whereas the numbers of vertices increase strongly with increasing r.

3.4.3 Stoichiometric robustness of networks

As was already indicated in the previous section, the total number of transitions and the numbers of mutations of different type may serve to investigate the robustness of network functions against exchanges of single reactions. We define a network to be *strongly robust* against a mutation if it maintains all its previous functions. A network is called *weakly robust* against a mutation if it loses some or all of its previous functions but remains functional.

We define the mean strong robustness of all networks of a given size r by the fraction of N- and G-mutations (which maintain all network functions) among all transitions

$$\rho_r^{\text{strong}} = \frac{T_r^{\text{N}} + T_r^{\text{G}}}{T_r^{\text{tot}}}. \tag{3.19}$$

Analogously, the mean weak robustness is defined by the fraction of PL-, CC-, and PC-mutations

$$\rho_r^{\text{weak}} = \frac{T_r^{\text{PL}} + T_r^{\text{CC}} + T_r^{\text{PC}}}{T_r^{\text{tot}}}. \tag{3.20}$$

Fig. 3.15 shows ρ_r^{strong} and ρ_r^{weak} as a function of the network size r. It is seen that for all possible values of r the relation $\rho_r^{\text{strong}} < \rho_r^{\text{weak}}$ holds true indicating that an arbitrary mutation is more likely to alter a network's function than to maintain all its functions. The general tendency is that the strong as well as the weak robustness increases with increasing r. This means that an arbitrary exchange of one reaction is more likely to result in an elementary network for larger network size. Obviously, this corresponds to the result given in Table 3.3 that the share of elementary networks among all networks

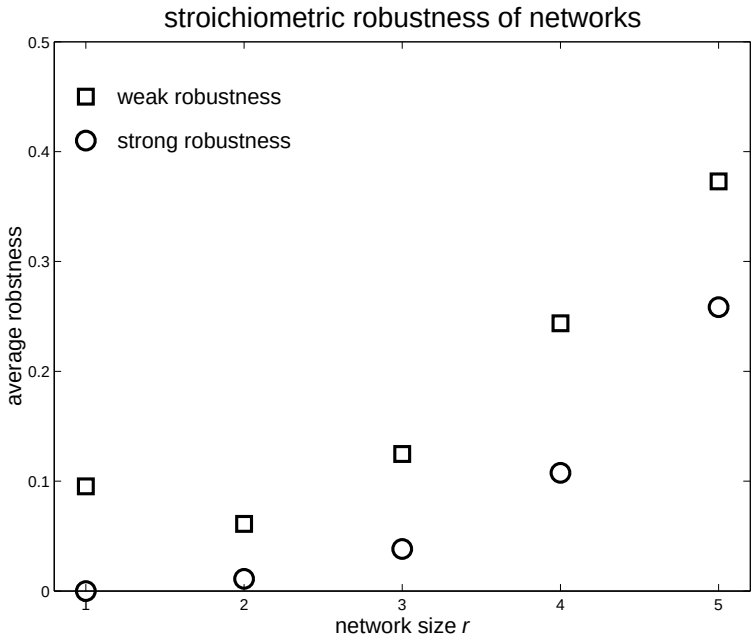


Figure 3.15: Robustness of carbon skeleton reaction networks. Show are the mean strong robustness and the mean weak robustness as a function of the network size r.

increases with r. The only exception is that $\rho_1^{\text{weak}} > \rho_2^{\text{weak}}$. For networks consisting of five reactions we get $\rho_5^{\text{weak}} + \rho_5^{\text{strong}} = 0.631$. Since the vast majority of elementary networks is of size $r = 5$ ($N_5 = 26334$, $N_1 + N_2 + N_3 + N_4 = 9108$) one may conclude that carbon skeleton reaction networks are very robust against mutational changes.

3.4.4 Islands of networks

Networks of size r with the ability to perform a given function $\langle a, b\rangle$ form a subgraph of G_r which will be denoted by $G_r(a, b)$. Analysis reveals that for $r = 3$ and $r = 4$ none of these subgraphs are connected in contrast to the complete graphs G_3 and G_4. The subgraphs consist of connected components of varying size which in the following will be called "islands". Networks of the same function but belonging to different islands can only be transformed into each other by a number of mutations forming a path on which the original function is temporarily lost.

For illustration, Fig. 3.16 shows the two subgraphs $G_3(6, 1)$ and $G_3(6, 2)$ embedded in the complete graph G_3. These two subgraphs consist of two and eleven islands, respectively. They have one complete island in common whose vertices are marked by

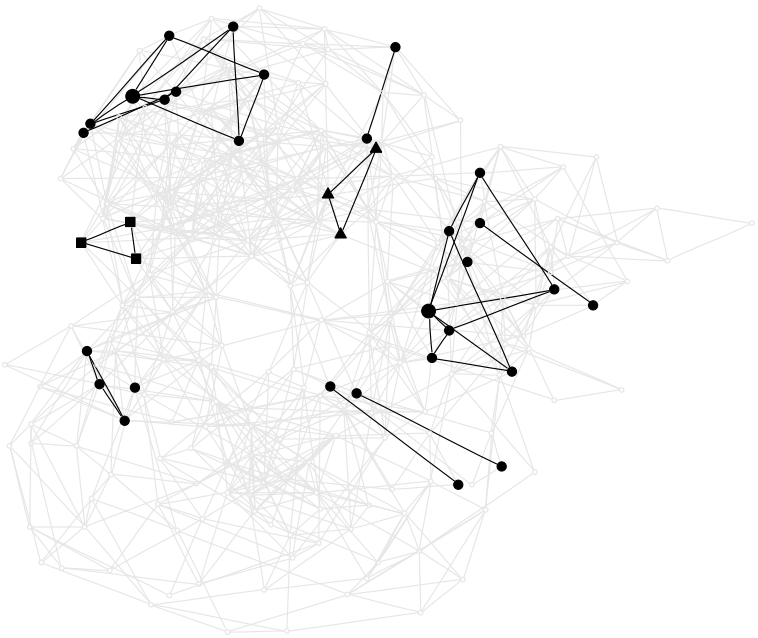


Figure 3.16: Decomposition of the subgraphs $G_3(6,1)$ and $G_3(6,2)$ into islands. Vertices representing elementary networks with the function $\langle 6,1\rangle$ are characterised by triangles and squares and networks with the function $\langle 6,2\rangle$ by circles and squares. The number and size of these islands correspond to the numbers given in the columns of Table 3.11 for the corresponding conversions. One island (squares) is multifunctional, i. e. all its networks belong to both subgraphs. The two islands in $G_3(6,2)$ with seven vertices are "highly symmetric", i. e. there exists in both cases a central vertex (larger circle) with edges to all other vertices, each of which is directly connected with two other neighbours.

squares. This is a "multifunctional island" in the sense that all networks included may perform both functions $\langle 6,1\rangle$ and $\langle 6,2\rangle$. In addition to that the subgraph $G_3(6,1)$ contains a second island of three networks (triangles). Besides the multifunctional island the subgraph $G_3(6,2)$ consists of ten more islands (circles) which are of different size (for details, see legend to Fig. 3.16). Interestingly, the islands are highly symmetric: the islands with three vertices have a "triangular" shape and the islands with seven vertices show a "hexagonal" shape, each having one central network which is neighboured to all six other networks of its island. These central networks are marked by slightly larger circles.

Table 3.11 gives an overview of the decomposition of all graphs $G_3(a,b)$ into islands.

Shown are the numbers of islands for a given size depending on the network function.

island size	function															
	$\langle 6,5\rangle$	$\langle 6,4\rangle$	$\langle 6,3\rangle$	$\langle 6,2\rangle$	$\langle 6,1\rangle$	$\langle 5,4\rangle$	$\langle 5,3\rangle$	$\langle 5,2\rangle$	$\langle 5,1\rangle$	$\langle 4,3\rangle$	$\langle 4,2\rangle$	$\langle 4,1\rangle$	$\langle 3,2\rangle$	$\langle 3,1\rangle$	$\langle 2,1\rangle$
1		2	25	2						2	19	4	2		13
2		5		5									2	2	
3	2	2		2	2	4	2	2	2	2		1	1	2	
5			2							2	1		2		1
7		2		2									2	2	
8						1	2	2	1		2				3
12													1		
17														1	
24										1		1			

Table 3.11: Decomposition of the graphs $G_3(a,b)$ into islands for all functions $\langle a,b\rangle$. Given are the numbers of islands depending on their size. Islands sizes occurring in the subgraphs $G_3(a,b)$ are given in the first column. Islands with other sizes do not occur.

Inspection of the Table 3.11 reveals that maximally 24 networks may occur in one and the same island and that there are gaps in the possible sizes of the islands (only 9 island sizes occur from islands consisting of only one network to islands of maximal size). The total number of islands of size one, i. e. of isolated networks (see the first row in Table 3.11) is much higher than the total number of other islands with a larger size. The majority of isolated networks perform one of the functions $\langle 2,1\rangle$, $\langle 4,2\rangle$, or $\langle 6,3\rangle$ which can already be performed by a single reaction. All other non-zero entries in the first row correspond to functions which can already be performed by two reactions (compare to Table 3.2). The majority of the graphs $G_3(a,b)$ contain islands of size three.

A thorough analysis shows that there is a symmetry in the structure of the islands. The symmetrical topology of the island of $G_3(6,2)$ with seven networks has been mentioned above (see Fig. 3.16). Indeed, all islands of the graphs $G_3(a,b)$ that consist of seven network possess the same wheel-like structure, i. e. the subgraphs defined by these islands are isomorphic. Islands of size five of the graphs $G_3(a,b)$ do also have a common structure, which can loosely be described as "half a wheel" – see Fig. 3.17

for clarification. Moreover, also the islands of size three of the graphs $G_3(a, b)$ are all

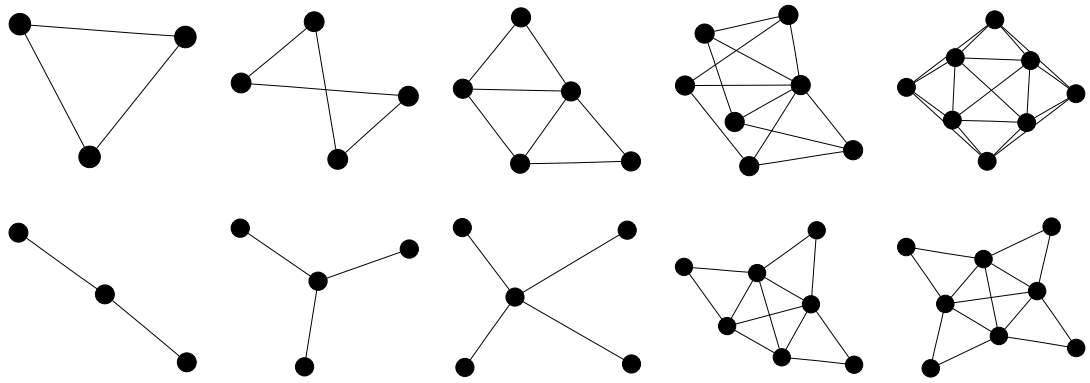


Figure 3.17: Topologies for small islands. For the islands sizes three, four, five, seven and eight (from left to right) two occurring topologies are displayed. The lower shapes only occur in the subgraphs $G_4(a, b)$ with the exception of islands of size eight, where both topologies are found in the subgraphs $G_3(a, b)$. For the other sizes, only the upper designs are found in the latter subgraphs.

triangular shaped as depicted in Fig. 3.16. However, there are two non-isomorphic designs for islands with eight networks which are also displayed in Fig. 3.17.

Note that Table 3.11 shows the same symmetry as Table 3.2, i. e. the columns for the conversions $\langle 6, 1\rangle$ and $\langle 6, 5\rangle$ are exactly the same as well as the columns for the conversions $\langle 6, 2\rangle$ and $\langle 6, 4\rangle$. Of course, this is no coincidence as can be seen as follows: In appendix B.2 a rule has been defined to map networks with the function $\langle L, a\rangle$ to networks with the function $\langle L, L - a\rangle$ and it becomes immediately clear from the definition of the map that neighbourhood relations are conserved. Consequently, the corresponding graphs must have the same topological structure. For the investigated case of $L = 6$ it means that the graphs $G_r(6, 5)$ and $G_r(6, 1)$ as well as the graphs $G_r(6, 4)$ and $G_r(6, 2)$ have the same composition of islands.

The compositions of the graphs $G_4(a, b)$ are presented in Table 3.12 which is organised in a different manner as Table 3.11 because there exist many larger islands of different sizes. Comparison of the two tables reveals that also for networks of size $r = 4$ there exists a large number of islands of size one, i. e. of isolated networks. And again, the majority of isolated networks perform one of the functions $\langle 2, 1\rangle$, $\langle 4, 2\rangle$, or $\langle 6, 3\rangle$ which can already be performed by a single reaction. Another common feature is that not all island sizes exist although the gaps in island sizes are not so clearly observed. However, the maximal island size for networks of size $r = 4$ is much larger (there exists one island with 354 networks). Further, even small islands (of size less

function	island size								
	1	2	3	4	5	6	7	8	>8
$\langle 6,5\rangle$,$\langle 6,1\rangle$	1	1					2		4 (56,60,75,171)
$\langle 6,4\rangle$,$\langle 6,2\rangle$		2					1		10 (13,16,21,23,29,36,39,64,77,109)
$\langle 6,3\rangle$	17				2				11 (4×13,20,6×31)
$\langle 5,4\rangle$	1	4		1			2		4 (41,45,123,354)
$\langle 5,3\rangle$		2		1			1		5 (16,43,45,144,317)
$\langle 5,2\rangle$	2	2	1	1			1		5 (16,43,45,144,317)
$\langle 5,1\rangle$		3	2						5 (16,43,56,60,354)
$\langle 4,3\rangle$	4	2	2				1		6 (10,45,64,79,96,211)
$\langle 4,2\rangle$	20			1				1	8 (13,21,29,29,31,31,41,82)
$\langle 4,1\rangle$	5	2	1		1		1		5 (38,60,79,85,240)
$\langle 3,2\rangle$	3	3	2		2				7 (10,13,35,37,96,109,140)
$\langle 3,1\rangle$		4			2				7 (12,13,45,47,70,109,173)
$\langle 2,1\rangle$	15					1			8 (13,20,29,29,31,40,41,82)

Table 3.12: Decomposition of the graphs $G_4(a,b)$ into islands for all functions $\langle a,b\rangle$. The numbers of small islands within each graph $G_4(a,b)$ are listed seperately up to islands of size eight. The numbers of larger islands are comprised into the last column where the island sizes are given as additional information in brackets.

than eight) differ in their topology. For island sizes up to eight, all shapes that occur are displayed in Fig. 3.17.

The interrelations between subgraphs of networks which may perform different functions may be of relevance in the context of the evolution of metabolism, in particular for the case that there exists a selection pressure to metabolise a new substrate. Evolutionary models which simulate such situations have been developed and are portrayed in section 3.6.

For completeness, also the subgraphs $G_5(a,b)$ for networks of size $r=5$ have been determined and astonishingly, they do not decompose into islands. Particularly this means that two networks of size $r=5$ performing the same function can always be interconverted into each other by consecutively exchanging a single reaction without having to give up the original function in an intermediate step.

3.5 Selected networks

In the previous sections (3.2–3.4) a thorough analysis has been presented concerning the complete set of possible network designs. Statistical results as well as properties of the graphs defined by the networks and mutations between networks have been presented. However, all these investigations considered the entirety of network structures rather than focusing on specific designs. In this section, some networks with outstanding properties are selected and will be presented to gain understanding of their structure.

3.5.1 Networks with the highest degree f of multifunctionality

As has been shown in section 3.2.2, networks of size $r = 2$ possess a maximal degree of multifunctionality $f = 3$. There are exactly two such networks, namely the network consisting of the two reactions $(0, 3|1, 2)$ and $(1, 3|2, 2)$ and the network consisting of the two reactions $(0, 6|2, 4)$ and $(2, 6|4, 4)$. These networks are labelled by "**K**" and "**F**", respectively, in Fig. 3.10. The first network can perform the three functions $\langle 3, 2\rangle$, $\langle 3, 1\rangle$ and $\langle 2, 1\rangle$ in a manner depicted in Fig. 3.18. The latter network (**F**) has exactly

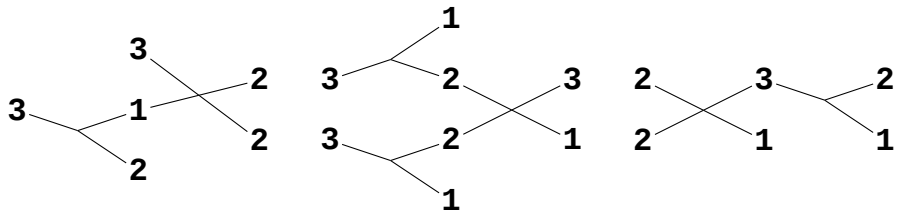


Figure 3.18: Network "**K**" performing the three functions $\langle 3, 2\rangle$, $\langle 3, 1\rangle$ and $\langle 2, 1\rangle$ (from left to right).

the same structure because it can formally be constructed from the first by doubling all numbers of carbon atoms in all participating compounds. Thus it can be immediately seen how the network can perform the functions $\langle 6, 4\rangle$, $\langle 6, 2\rangle$ and $\langle 4, 2\rangle$.

For networks of size $r = 3$, the investigations in section 3.2.2 revealed that the maximal degree of multifunctionality amounts to $f = 6$ and that there are five such networks. For illustration, the location of these five networks is visualised in Fig. 3.19. It is seen that three of these five networks are neighboured in the sense that they can be transformed into each other by exchange of a single reaction. These three networks all consist of three out the following set of four reactions:

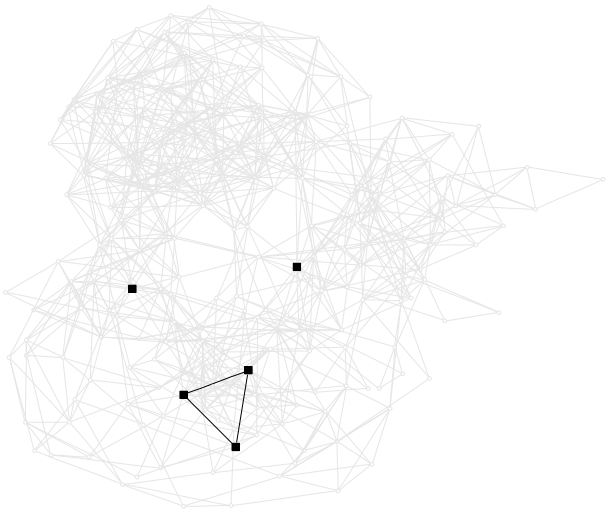


Figure 3.19: The location of the five networks within G_3 with maximal functionality $f = 6$ are marked by black squares. Edges representing mutations between these networks are highlighted.

$\{(1,5|3,3),(3,5|4,4),(0,4|1,3),(0,5|1,4)\}$, i. e. the reactions 4, 11, 17 and 18 from Table 3.1. One network comprises the first three reactions, another uses the first, third and fourth reaction and the third uses the first, second and fourth reaction of this set.

As an example, a closer look will be taken at the latter network, i. e. the network consisting of the three reactions $(1,5|3,3)$, $(3,5|4,4)$, and $(0,5|1,4)$. It may perform the six conversions $\langle 3,1\rangle$, $\langle 4,1\rangle$, $\langle 5,1\rangle$, $\langle 4,3\rangle$, $\langle 5,3\rangle$ and $\langle 5,4\rangle$ in an elementary way. Fig. 3.20 shows six different representations of this network corresponding to each of its functions. The stoichiometric matrix of this network is

$$\mathbf{N} = \begin{pmatrix} 0 & 0 & -1 \\ -1 & 0 & 1 \\ 0 & 0 & 0 \\ 2 & -1 & 0 \\ 0 & 2 & 1 \\ -1 & -1 & -1 \\ 0 & 0 & 0 \end{pmatrix} \tag{3.21}$$

and the solutions $V_{\langle a,b\rangle}$ of Eq. (3.2) corresponding to the six functions are

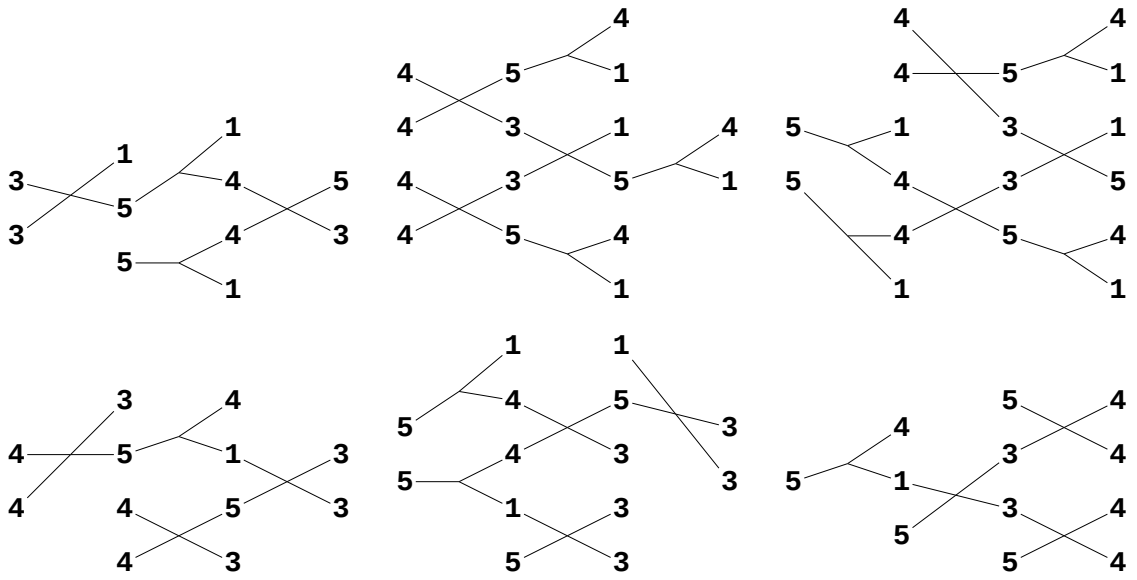


Figure 3.20: The network of size three consisting of the reactions $(1,5|3,3)$, $(3,5|4,4)$, and $(0,5|1,4)$. This network has the maximal degree of multifunctionality $f = 6$. The six functions it can perform are visualised. They are from top left: $\langle 3,1\rangle$, $\langle 4,1\rangle$, $\langle 5,1\rangle$, $\langle 4,3\rangle$, $\langle 5,3\rangle$ and $\langle 5,4\rangle$.

$$
\begin{aligned}
V_{\langle 3,1\rangle} &= \begin{pmatrix} -1 \\ -1 \\ 2 \end{pmatrix}, \quad V_{\langle 4,1\rangle} = \begin{pmatrix} -1 \\ -2 \\ 3 \end{pmatrix}, \quad V_{\langle 5,1\rangle} = \begin{pmatrix} -1 \\ -2 \\ 4 \end{pmatrix}, \\
V_{\langle 4,3\rangle} &= \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}, \quad V_{\langle 5,3\rangle} = \begin{pmatrix} 2 \\ -1 \\ 2 \end{pmatrix}, \quad V_{\langle 5,3\rangle} = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}.
\end{aligned}
\tag{3.22}
$$

3.5.2 The largest distance within G_3

The investigations presented in section 3.4.2 yielded that the diameter for the graph G_3 amounts to $D(G_3) = 8$. A closer inspection shows that there are exactly five pairs of networks which have this maximal mutual distance. Further, one finds that these five pairs have one network in common. Figuratively spoken, this network can be said to be located at one end of the graph G_3 whereas the other five networks lie on the opposite end. This fact is illustrated by Fig. 3.21. The network on the right-hand side

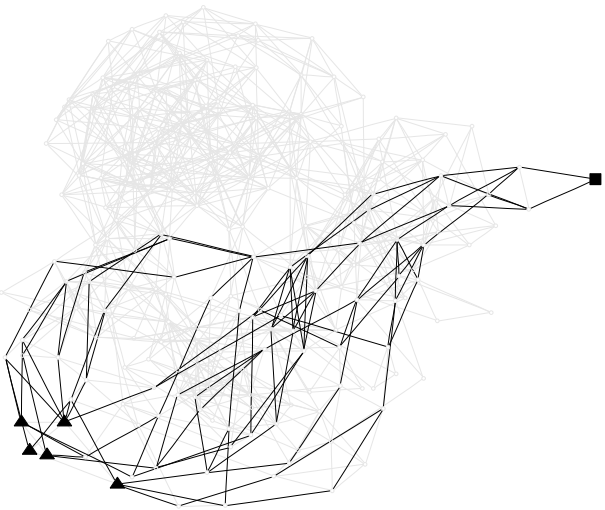


Figure 3.21: Shown are the networks with the maximal distance from each other. The five networks marked by triangles all have the distance $d = 8$ from the network marked by the square, the "most remote" network. The networks marked by triangles can result from the network marked by the square in a sequence of at least eight mutations. The edges along all possible paths of length eight are coloured black.

marked by the square can therefore be called the "most remote" network. It consists of the reactions $(3, 5|4, 4)$, $(3, 6|4, 5)$ and $(0, 6|3, 3)$ and has the functions $\langle 6, 5\rangle$, $\langle 6, 4\rangle$, $\langle 5, 4\rangle$, $\langle 5, 3\rangle$ and $\langle 4, 3\rangle$.

3.5.3 Central networks in G_3

Together with the diameter of the graph G_3, the average distance between networks has been calculated in section 3.4.2 to be $\bar{d}(G_3) = 3.55$. Further calculation was used to identify the network which has the shortest mean distance to all other networks. This network's location within G_3 is visualised in Fig. 3.22 as the vertex marked by a square. The mean distance from this network to all other networks amounts to 2.83 which is considerably lower than the overall average distance $\bar{d}(G_3)$. This network consists of the reactions $(0, 2|1, 1)$, $(0, 3|1, 2)$ and $(0, 4|2, 2)$ and is monofunctional with $\langle 4, 3\rangle$ being the only conversion it can perform in an elementary way – see the left diagram in Fig. 3.23. It possesses 18 neighbours, which is the second largest number of neighbours found for networks within G_3 – see Fig. 3.12. The only network with a larger number of neighbours possess 22 neighbours and the corresponding vertex in

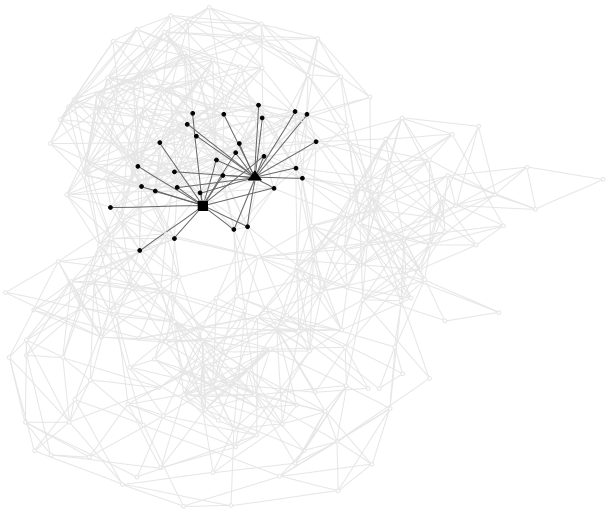


Figure 3.22: The central networks in G_3. The network with the smallest mean distance to all other networks in G_3 is marked by a square, the network with the largest number of neighbours (22) is marked by a triangle. In order to demonstrate the high connectivity of the vertices representing these networks, all possible mutations from both networks have additionally been emphasised by black edges.

Fig. 3.23 is marked by a triangle. This is also a monofunctional network which performs the function $\langle 3, 1\rangle$ using the three reactions $(0, 2|1, 1)$, $(0, 4|2, 2)$ and $(3, 3|2, 4)$ – see the right diagram in Fig. 3.23. The high connectivities of these networks has been visualised by marking all edges representing possible mutations originating from one of these networks.

The network with the highest number of neighbours is the most robust network in G_3 in the sense that a random mutation is least likely to result in a non-functional network. A closer inspection reveals that eight from the 22 neighboured networks can also perform the function $\langle 3, 1\rangle$, whereas the remaining 14 networks perform exclusively functions other than $\langle 3, 1\rangle$. Taking into account that 57 transitions – see Eq. (3.12) – originate from this network, the values for the strong and weak robustness amount to 0.140 and 0.246, respectively. These values are significantly larger than the mean values for $r = 3$ – see Fig. 3.15.

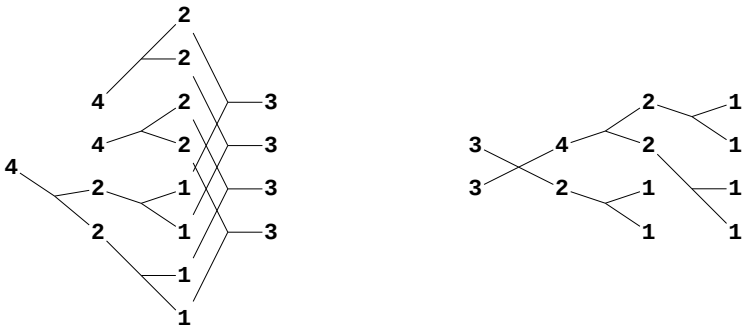


Figure 3.23: The two central networks from Fig. 3.22 performing their only functions $\langle 4, 3\rangle$ and $\langle 3, 1\rangle$, respectively.

3.5.4 The completely robust network

In section 3.4.2 it was mentioned that there exists exactly one network with the property that an arbitrary exchange of one of its reactions will result in another functional network. This network is the only network with the highest theoretically possible connectivity given in Eq. 3.12. It is a network of size $r = 5$ with 85 neighbours and consists of the five reactions $(3, 5|4, 4)$, $(0, 5|2, 3)$, $(1, 6|3, 4)$, $(0, 6|1, 5)$ and $(1, 3|2, 2)$. It can perform 13 conversions in an elementary way and the remaining two in a non-elementary way. As the diagrams for the network performing the elementary functions become quite large, only one example has been selected. The two non-elementary conversion are performed by this network using only three out of its five reactions. Fig. 3.24 shows how the network can perform the conversion $\langle 3, 2\rangle$ in an elementary way (left side) and how the conversions $\langle 6, 4\rangle$ and $\langle 6, 2\rangle$ are performed in a non-elementary way.

This remarkable feature might be of biological relevance since regardless which reaction is affected by a mutation, the network still remains functional and therefore the damage is limited. It is the most robust network of all elementary networks. Since this network already performs a large number of functions ($f = 13$), it is not surprising that an arbitrary exchange of one reaction results in the loss of at least some of the network's original functions. Indeed, this is the case for all mutations originating from this network. Therefore the values of the strong and weak robustness amount to $\rho^{\text{strong}} = 0$ and $\rho^{\text{weak}} = 1$. Detailed analysis shows that 54 of the 85 possible mutations originating from this network result in a partial change of functions (PC-mutations), 28 mutations result in a partial loss of the network's funcitons (PL-mutations) and three mutations involve a complete change of functions (CC-mutations).

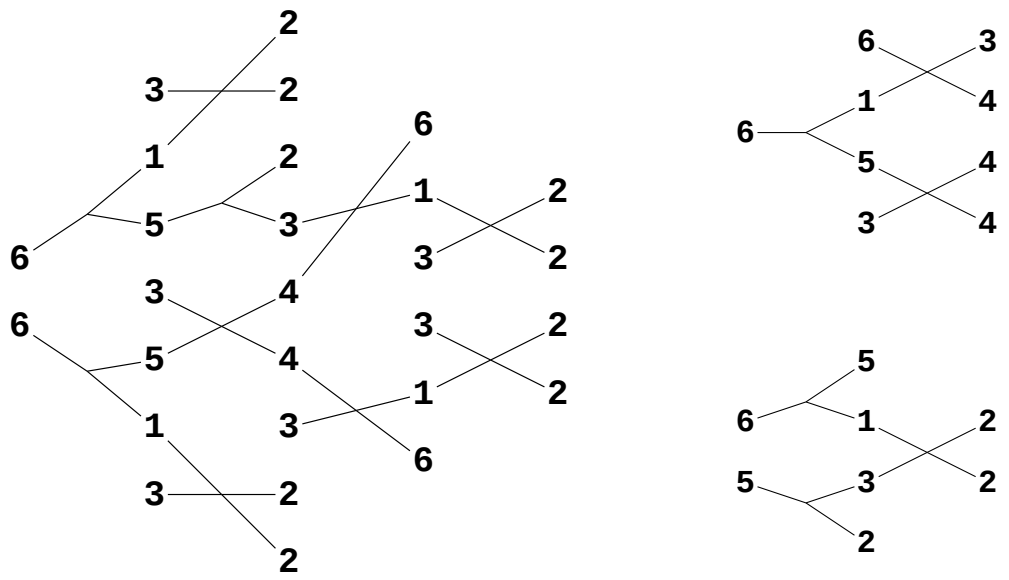


Figure 3.24: The completely robust network performing the function $\langle 3, 2\rangle$ and the (non-elementary) conversions $\langle 6, 4\rangle$ and $\langle 6, 2\rangle$.

3.5.5 Bi-uni-networks and bi-bi-networks

The analysis presented so far in this chapter takes into account two types of reactions, i. e. bi-bi-reactions and bi-uni-reactions (see Table 3.1). For both types of reactions there exist many examples of enzyme catalysed processes taking place in cellular metabolism (see section 3.1.2). Bi-bi-reactions often occur as overall enzyme catalysed reactions. Considering non-enzymatic reactions or elementary steps in complex enzymatic reactions, those reactions which are bimolecular in both directions may be inappropriate as building blocks. In fact, every such reaction can be considered to be a combination of two bi-uni reactions. A possible decomposition of a reaction $C_i + C_j \rightleftharpoons C_k + C_l$ into two "hemi-reactions" reads: $C_i + C_{l-i} \rightleftharpoons C_l, C_j \rightleftharpoons C_{l-i} + C_k$, where without loss of generality we have chosen $l > i$. Since the relation $i + j = k + l$ holds true, both hemi-reactions fulfil the condition that the number of carbon atoms is conserved by the reaction. Note that the number of carbon atoms of the newly introduced compound C_{l-i} is less than the maximal number of carbon atoms of the original reactants. For the case $L = 6$ one may conclude that every reaction 1–13 of Table 3.1 can be viewed as a superposition of two bi-uni-reactions (reactions 14–22 in Table 3.1). It may be therefore of interest to investigate sets of "bi-uni-networks" consisting exclusively of the latter type of reactions. We have performed such an analysis

by applying the methods described above. Some of the results can be summarised as follows:

1. The total number of elementary bi-uni-networks is significantly smaller than in the former case, i. e. : $r = 2$: 3 (18), $r = 3$: 14 (213), $r = 4$: 42 (2160) and $r = 5$: 81 (14152), where the numbers in brackets indicate the numbers of networks for the original set of 22 reactions. The networks consisting of the limited set of reactions form a subset of all networks originally constructed. For the case $r = 3$ all networks of this subset have been identified and are represented by black circles in Fig. 3.25.

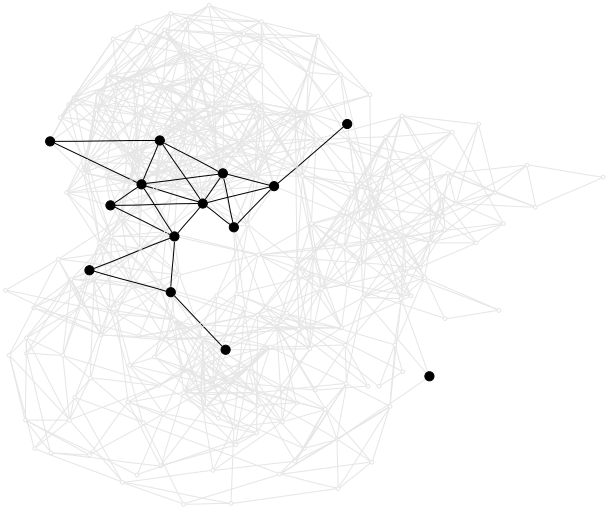


Figure 3.25: All networks of size $r = 3$ consisting entirely of bi-uni-reactions. The black edges represent all possible mutations between bi-uni-networks. Clearly, an isolated network can be made out while the other 13 form a connected subgraph.

2. As for networks assembled from the complete set of reactions, there exist among the bi-uni-networks monofunctional and multifunctional networks. Networks of size $r \leq 3$ cannot perform all conversions in an elementary way (networks of size $r = 2$ cover five functions, networks of size $r = 3$ cover 13 functions). For $r > 3$ elementary networks exist for all possible conversions.

3. The graphs with these networks of a fixed size as vertices and possible mutations as edges are all connected except for the case $r = 3$. As can be seen in Fig. 3.25, in

this case there exists one isolated network whereas the other 13 form a connected subgraph.

The next question that arises is whether there exist elementary networks consisting exclusively of bi-bi-reactions and which may interconvert two external compounds C_a and C_b. In Appendix B.4 it is proven that such bi-bi-networks do not exist.

3.5.6 Glycolysis and the Citric Acid Cycle

The presented model includes the description of some important metabolic systems found in living cells. The central energy metabolism in aerobic organisms comprises the glycolytic pathway combined with the citric acid (TCA) cycle. Considering only the changes in the numbers of carbon atoms in the participating compounds, this metabolic system is in the framework of this model described as a network with $r = 5$ reactions consisting of the reactions $(0, 6|3, 3)$, $(0, 3|1, 2)$, $(0, 6|2, 4)$, $(0, 6|1, 5)$ and $(0, 5|1, 4)$. For the detailed correspondence between these reactions and their real existing counterparts see section 3.1.2.

All these reactions are bi-uni-reactions which makes the network a bi-uni-network. This by itself is an outstanding property since only 81 out of all $Q_5 = 14152$ networks of size $r = 5$ are bi-uni-networks (see section 3.5.5).

As all elementary networks with five reactions, this network can perform all conversions $\langle a, b\rangle$. It can perform ten functions in an elementary way, whereas the five conversions $\langle 6, 3\rangle$, $\langle 4, 2\rangle$, $\langle 3, 2\rangle$, $\langle 3, 1\rangle$ and $\langle 2, 1\rangle$ can be performed in an non-elementary way, i. e. using only a subset of its five reactions.

The main purpose of the central energy metabolism is to oxidise glucose (a hexose) into carbon dioxide and store the energy which is won during this process in form of ATP molecules (see the detailed analysis in chapter 2). This conversion $\langle 6, 1\rangle$ is one of the network's functions and its graphical representation is given in Fig. 3.26. The cyclic structure becomes clear by keeping in mind that the loose ends correspond to the same compound C_5. A hexose is split into two trioses which are decarboxylated to yield a one- and a two-carbon structure (CO_2 and an aldehyde group, respectively). The two-carbon structure enters the cycle by being combined with a four carbon molecule to form a C_6-compound from which two C_1-compounds are removed in consecutive steps.

The network possesses 50 neighbours within G_5, 13 of which are also bi-uni-networks. A closer inspection shows that among these 50 mutations there are six N-mutations and seven G-mutations. Since these mutations maintain all of the network's original

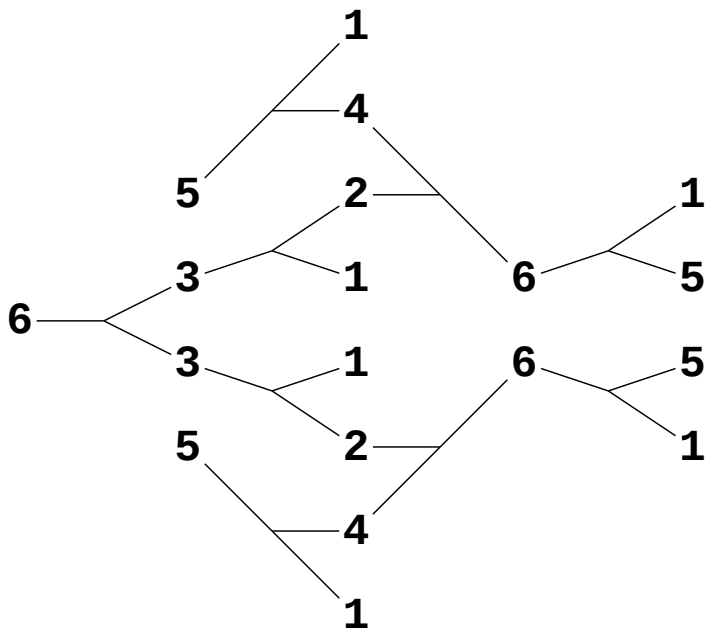


Figure 3.26: Graphical representation of the central energy metabolism as described by the model for the function $\langle 6, 1\rangle$ – decomposition of glucose (C_6) into carbon dioxide (C_1).

functions, the strong robustness for this network results to $\rho^{\text{strong}} = 0.153$. Further, there are 16 PC- and 21 PL-mutations. However, there is no CC-mutation originating from this network. Thus, the value for the weak robustness of this network is $\rho^{\text{weak}} = 0.435$. These values are not very outstanding with the value for the strong robustness being somewhat smaller than the average value for all networks of size $r = 5$ ($\rho_5^{\text{strong}} = 0.258$) and the value for the weak robustness being somewhat larger than the average ($\rho_5^{\text{weak}} = 0.373$). The fact that there are no CC-mutations originating from this network means that in all cases a mutation will result in a network which can perform at least one function of the original network.

If only bi-uni-reactions are considered as building blocks, the network has 13 neighbours. Here, seven of the corresponding mutations lead to a partial loss of functions (PL), four to a partial change of functions (PC) and two involve a gain of functions (G).

3.6 Evolutionary models

The networks considered in the previous sections are able to metabolise different kinds of substrates into different products and therefore possess realistic biological functions. For given external conditions characterised by the presence of one or more substrates C_a, only those networks may perform a conversion which can make use of these substrates. If the metabolism of a cell or a whole organism contains such a network, it has a higher fitness for the present environment resulting in a selective advantage. In the following, selection and mutation processes among the set of possible networks are considered in order to characterise the evolution towards networks with special advantageous properties.

The dynamics of evolutionary processes can be studied on the graphs G_r where the vertices represent networks and the edges possible mutations. In the following we apply evolutionary algorithms which are based on principles of stochastic processes (Ebeling et al. 1990). Specifically, the model considers a population of networks, rules for transitions between networks according to possible mutations, as well as reproduction and selection of networks according to their evaluation by a fitness function – for early work in this direction see Ebeling and Feistel (1977), Heinrich and Sonntag (1981).

The general functioning of the algorithm has been described in section 1.3. The methodology is very similar to the optimisation procedure in section 2.2 but the main focus is slightly different. Whereas in chapter 2 the algorithm was used to determine optimal reaction sequences, we are here more interested in the evolutionary process itself.

In the cases examined below the fitness function ϕ is defined in a very simple way, amounting to $\phi = 1$ for a network which possesses a desired function and $\phi = 0$ otherwise.

There exist two fundamentally different possibilities concerning the dependency of the networks and their environment. Either, the environmental conditions are independent from the processes taking place within the population, or the environmental conditions are affected by these processes. The first case may occur for unlimited resources while the second case is a realistic scenario for limited resources. An interesting aspect of the latter case is reintegration of the products of networks into the environment which then may be consumed by the population. In this section we confine our analysis to populations consisting of networks with $r = 3$ reactions.

3.6.1 Optimisation under imposed environmental conditions

In this scenario the development of a population is investigated under the condition that the environment provides certain substrates which are present in excess. The availability of these substrates may be permanent or time dependent. The evolutionary dynamics are analysed assuming permanent availability of one or two compounds or alternating availability of these two compounds.

Let us assume for this scenario that the environmental conditions are such that the compounds C_4 and C_6 are permanently available. It is assumed that those networks have a selective advantage which consume these compounds and perform the conversions $\langle 4, 1\rangle$ or $\langle 6, 1\rangle$. These networks are characterised by a fitness $\phi = 1$ and by a reproduction probability $q = 0.1$. According to Table 3.2 there are 31 networks of size $r = 3$ which can perform the function $\langle 4, 1\rangle$ and six networks which perform the function $\langle 6, 1\rangle$. Closer inspection shows that these two subsets have exactly one network in common, i. e. a network which is bifunctional with respect to these two conversions. Table 3.11 shows that the six networks with the function $\langle 6, 1\rangle$ form two islands each containing three networks. The 31 networks with the function $\langle 4, 1\rangle$ decompose into six islands, one very large island of size 24, one island with three networks and four isolated networks.

It is useful to illustrate this situation in order to gain insight into the topology of the set on which the evolutionary models are "living". Fig. 3.27 shows where the islands are located within the graph G_3. Vertices marked by circles represent networks with the function $\langle 4, 1\rangle$, vertices marked by triangles represent networks with the function $\langle 6, 1\rangle$. The one vertex which is marked by the square represents the only network which is bifunctional with respect to these two functions. This notation will be used in all graphs presented in this section. The following properties of the islands are noteworthy: One island of networks with the function $\langle 6, 1\rangle$ is neighboured to the large island (24 networks with the function $\langle 4, 1\rangle$) in the sense that it is possible to reach one island from the other by a single mutation. The other island of networks with the function $\langle 6, 1\rangle$ is separated from the large island, i. e. more than one mutation is necessary to convert networks from these two islands into another. One network of this island is identical to the bifunctional network, which – considering only the function $\langle 4, 1\rangle$ – is an isolated network. The remaining islands of networks with the function $\langle 4, 1\rangle$ (one with three networks and three isolated networks) are also separated from the other islands.

The initial population consists of $n = 100$ copies of one specific network which does

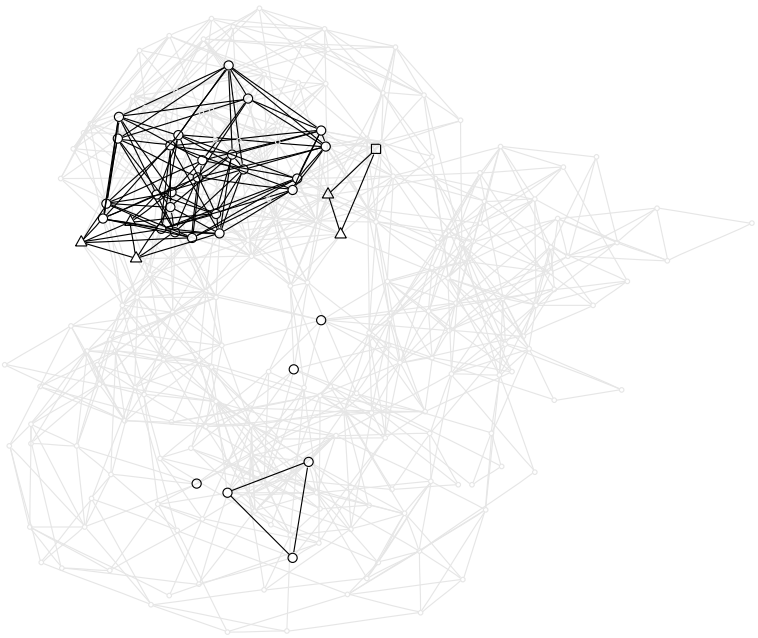


Figure 3.27: The location of the subgraphs $G_3(6,1)$ (triangles) and $G_3(4,1)$ (circles). The vertex marked by a square represents the network which can perform both functions $\langle 4,1\rangle$ and $\langle 6,1\rangle$.

not perform one of these two conversions. As the initial network we have chosen the "most remote" network which has largest mean distance to all other networks (see section 3.5.2 and Fig. 3.21 on page 76). The composition of the population will change in time due to mutation mechanisms. The probability that a network will be altered by a mutation during one generation is set to $p = 0.05$.

Fig. 3.28 gives a graphical representation of this evolutionary process over the first 400 generations on the graph G_3. The edges of each graph in Fig. 3.28 are weighed proportional to the frequency of the corresponding transitions counted over 100 generations where the four graphs represent the development over four consecutive periods of 100 generations. A darker edge indicates a higher weight (see legend to Fig. 3.28).

It is seen that the population has evolved towards a cluster of networks which can perform one of these functions. Networks with a high number of neighbours which perform one of the two favourable functions are most abundant. Other networks which have also a selective advantage but are remote of this cluster are not generated to a significant amount. Fig. 3.28a reveals traces of the evolutionary process. The dark edges around the initial networks are the remnants of the early stages of the process

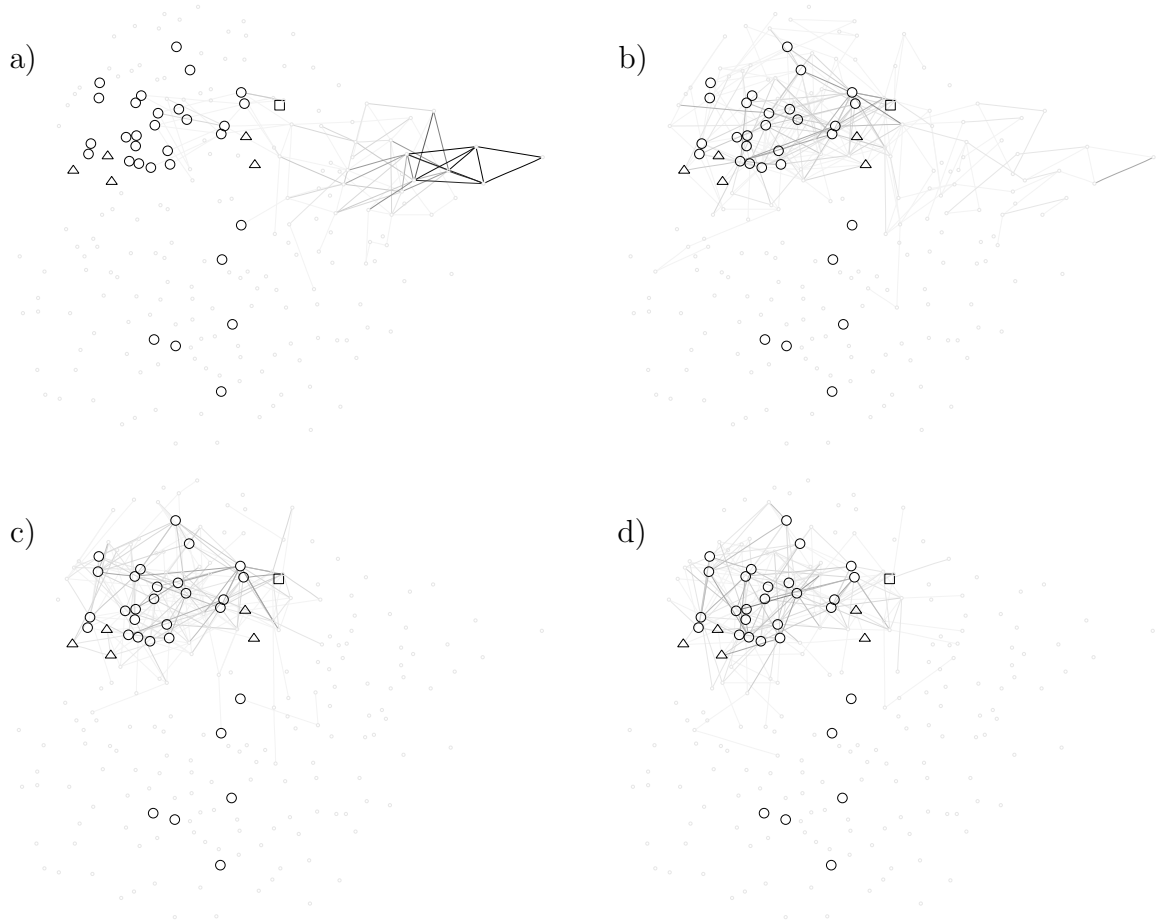


Figure 3.28: Graphical representation of an evolutionary process where the networks with the functions $\langle 4, 1\rangle$ and $\langle 6, 1\rangle$ have a selective advantage. Shown are four periods over 100 generations each (parts a–d). Darker edges indicate a higher frequency of the represented mutation.

where the system performs a random search among networks without any selective advantage. In the generations 101–200 (Fig. 3.28b), transitions which occurred during intermediate stages of the development are clearly seen. In the next two periods of 100 generations (Figs. 3.28c,d) almost exclusively such transitions are observed which originate from one network with a selective advantage. It can be concluded that the population has stabilised at this stage.

The process has been continued up to generation 2000 and it was observed that the population qualitatively did not change considerably.

Starting from the population which evolved after these 2000 generation, continuations with changed environmental conditions have been investigated. For the continuing

simulations a lower mutation rate of $p = 0.01$ was chosen. In particular, the two cases have been considered in which not two but only one resource (C_4 or C_6, respectively) remains available. Fig. 3.29 shows a graphical representation of the process in which

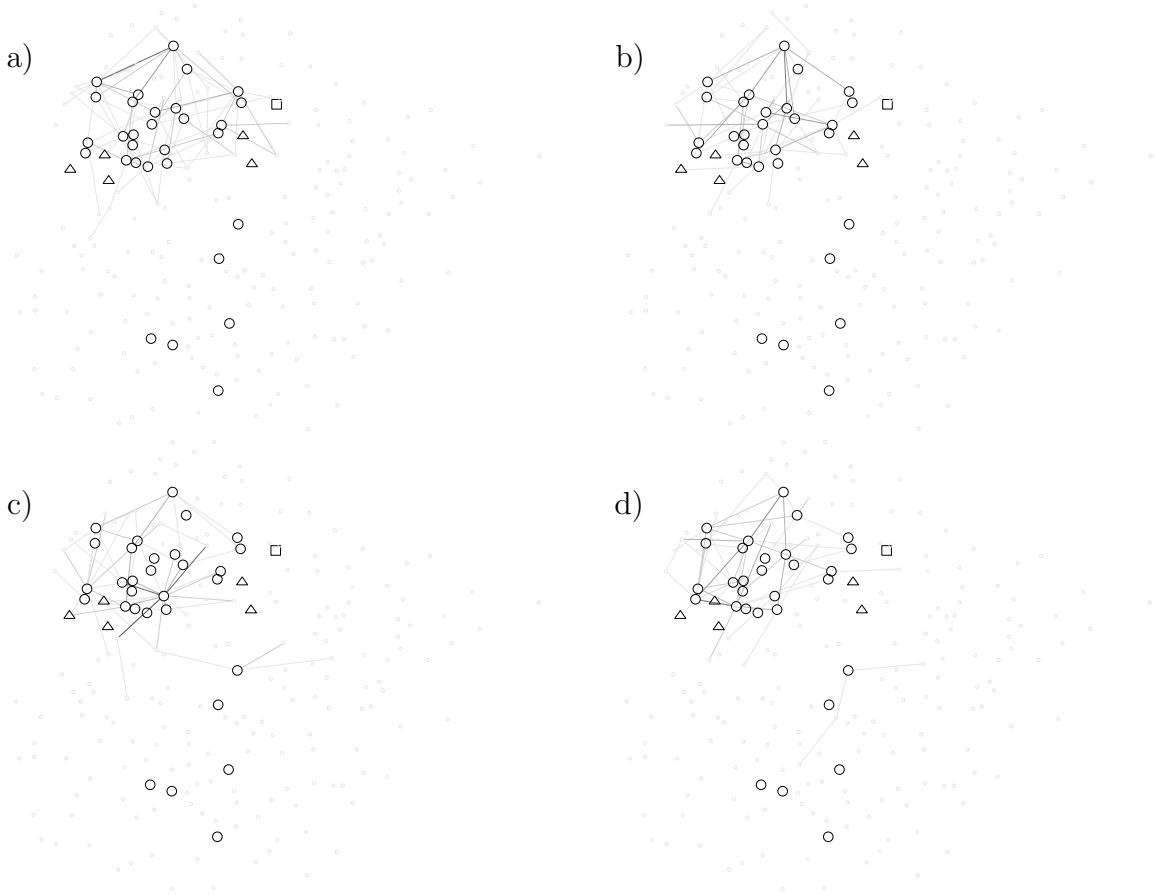


Figure 3.29: Graphical representation of an evolutionary process where the networks with the function $\langle 4, 1\rangle$ have a selective advantage. The first two periods of 100 generations are represented in parts a and b. Generations 401–600 are represented parts c and d.

only those networks possessing the function $\langle 4, 1\rangle$ have a selective advantage whereas Fig. 3.30 represents the case with the only available resource being C_6. As expected, both figures reveal that for the new environmental conditions those networks become dominant which can make use of the present resource by performing the corresponding conversion. However, not all networks having a selective advantage are present to a significant amount.

In the case of the C_4-resource (Fig. 3.29), a cluster of networks survives which corresponds to the island of the subgraph $G_3(4, 1)$ containing 24 networks (see Table 3.11).

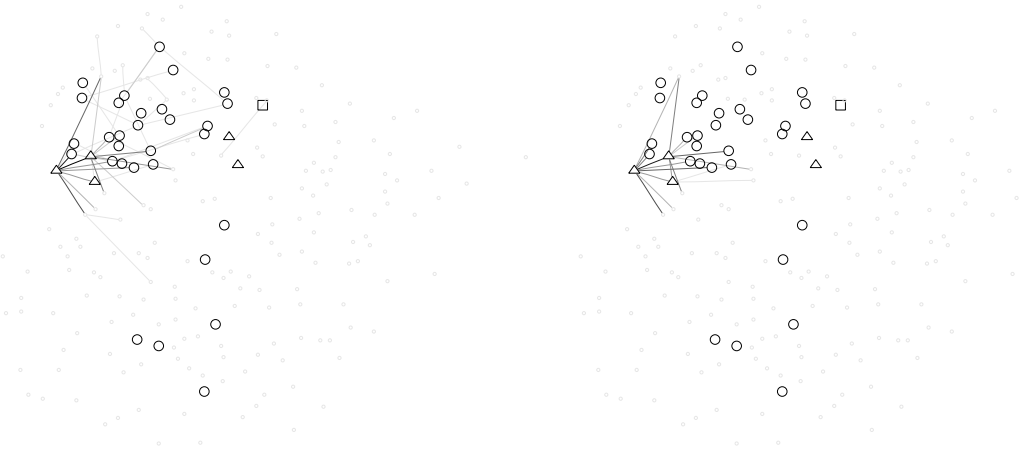


Figure 3.30: Graphical representation of an evolutionary process where the networks with the function $\langle 6, 1\rangle$ have a selective advantage. The first two periods of 100 generations are represented.

The other seven networks with the function $\langle 4, 1\rangle$ (including the bifunctional network) belonging to the other islands occur only sporadically. That they do indeed occur can be seen by following the process for a longer time. Figs. 3.29c,d show the generations 401–600 of this process. Here, the temporary occurrence of an isolated network can clearly be seen (near the centre of the graph).

In the case of the C_6-resource (Fig. 3.30) the population evolves towards networks belonging to one of the two islands of the subgraph $G_3(6, 1)$, see Table 3.11. This island was favoured compared to the other one since its networks and neighboured networks were more abundant when the environmental condition changed. However, due the randomness of selection and mutation a repetition of the process could result with some probability in the dominance of the other island.

As becomes apparent from Figs. 3.29 and 3.30, the population evolves in each case towards the networks having a selective advantage under the imposed environmental conditions and their neighbours. This means, networks persisting over a longer period of time form a "quasi-species" (Eigen 1971), i. e. they are closely related in a sense that they can be interconverted by a small number of mutations.

For quantifying the mean distances of the networks involved in a given stage of the evolutionary process, the extension of a population can be defined in the following way:

$$\bar{d}_P = \frac{1}{n(n-1)} \sum_{i,j}^{Q_r} n_i n_j d(\mathcal{N}_i, \mathcal{N}_j) \tag{3.23}$$

In this definition the distances between networks are weighed according to their occurrence n_i. For the special case $n_i = 1$ for all i, representing a uniform distribution of networks where every network $\mathcal{N}_i$ is present in exactly one copy, the relation $n = Q_r$ holds such that the extension $\bar{d}_P$ equals the mean distance of the networks within the graph G_r. A population is characterised by $\bar{d}_P = 0$ if it consists of copies of only one network $\mathcal{N}_i$.

Fig. 3.31 presents time courses of the extension of the populations for the evolutionary processes depicted in Figs. 3.28–3.30. They start from $\bar{d}_P = 0$ since the initial

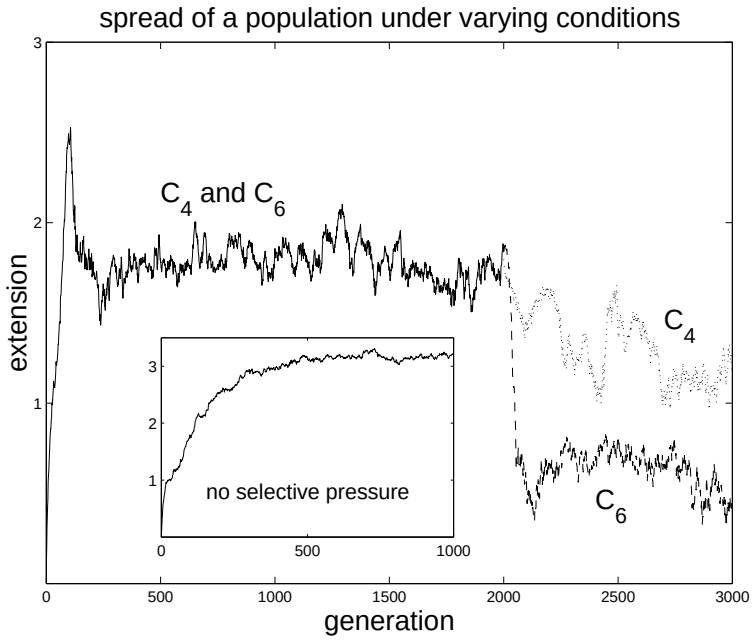


Figure 3.31: Development of the extension of a population with two resources available (up to generation 2000), and only one resource remaining present (after generation 2000). The inset compares a process without selective pressure.

population consists of n copies of one special network. The environmental conditions are such that until generation 2000 the compounds C_4 and C_6 are permanently available. The evolution starts with a search phase characterised by a sharp increase of the extension of the population. Around generation 100 a favourable network is found (in this case a network with the function $\langle 4, 1\rangle$). Due to its higher reproduction rate this network acts as a seed for establishing a quasi-species. Accordingly, the next phase is characterised by a rapid increase in the number of the copies of this favourable network

and its close neighbours resulting in a sharp decline of the extension of the population. It soon levels off during the generations 150–250 at a rather high value for $\bar{d}_P$ about 1.8 ± 0.15 when the quasi-species has been established. The reason that the extension of the population does not further decrease is that the quasi-species consists not only of one "master species" and its mutants but of a cluster of equally favoured networks and their mutants. However, this value is significantly lower than the mean distance $\bar{d} = 3.55$ between all pairs of networks of G_3. Until generation 2000 the process is characterised by the persistence of the established quasi-species which becomes apparent by slight fluctuations of $\bar{d}_P$.

Starting from generation 2000, two time courses are shown in Fig. 3.31 representing the development of the extensions after a switch of the external conditions from the simultaneous presence of C_4 and C_6 to the presence of only one of these two compounds. The two curves correspond to the simulations depicted in Figs. 3.29 and 3.30. In both cases the time courses are characterised by a decline of the extensions which is partly explained by the fact that the mutation probability decreased by a factor of five. The decline of the population extension in the case of C_6-resources is very pronounced since only three out of six favourable networks act as seeds for establishing the new quasi-species (see Fig. 3.30). In the case of C_4-resources the quasi-species contains 24 networks resulting in a moderate decline of the extension. These processes differ considerably from a process without selective pressure for which there is a permanent increase in the extension of the population (see inset in Fig. 3.31).

The evolutionary algorithm can also be applied to environmental conditions changing in time. In the following it is assumed that the two resources, C_4 and C_6 are alternately present, i. e. there exist periods in which networks with either the function $\langle 4, 1\rangle$ or the function $\langle 6, 1\rangle$ have a selective advantage. Fig. 3.32 shows the development of the population when the available resources alternates every 100 generations starting with the resource C_4. Shown are percentages of networks in the population with either the function $\langle 6, 1\rangle$ (but not $\langle 4, 1\rangle$), or the function $\langle 4, 1\rangle$ (but not $\langle 6, 1\rangle$), as well as the percentage of the one network which is bifunctional with respect to these conversions. Again, the process starts from the population which emerged after continuation of the process depicted in Fig. 3.28 for 2000 generations. It was continued for another 2000 generations with a lower mutation rate of $p = 0.01$. Initially, the periodic changes in the environment are reflected by periodic changes in the composition of the population. In the present case this holds true for five cycles each consisting of two different periods. Around generation 1000, suddenly a significant amount of the bifunctional network appears in the population and dominates after a short transition time. Since

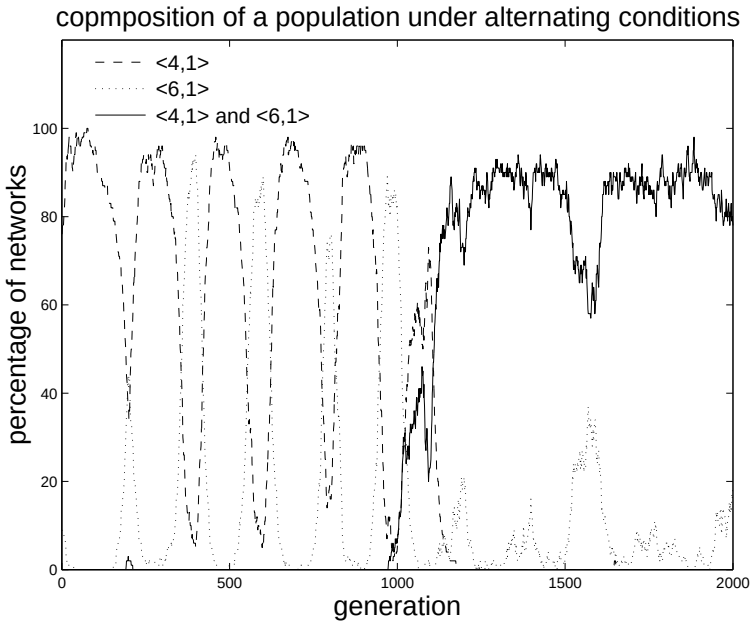


Figure 3.32: The percentages of networks able to perform one function (either $\langle 4,1\rangle$ or $\langle 6,1\rangle$) and both functions within the population. The availability of these two resources changes every 100 generations.

this network has a selective advantage under both environmental conditions, oscillations in the composition of the population become less significant, although still visible for networks with only one of the two functions which are present in low numbers.

It is interesting to examine the emergence of the bifunctional network in more detail. Fig. 3.33 shows the transitional phase from an oscillating behaviour of the composition of the population (see Fig. 3.32) to the dominance of the bifunctional network. Every graph shows which mutations were performed during a period of 100 generations. Again, darker edges indicate a higher frequency. Shown are six periods of 100 generations starting with generation 701 in Fig. 3.33a. Here, the bifunctional network has not yet prevailed. In Fig. 3.33c (generations 901–1000) the occurrence of mutations originating from / ending in the bifunctional network can be made out. In Fig. 3.33d (generations 1001–1100), these mutations become more pronounced until in Figs. 3.33e,f (generations 1101–1300) the bifunctional network clearly dominates the population.

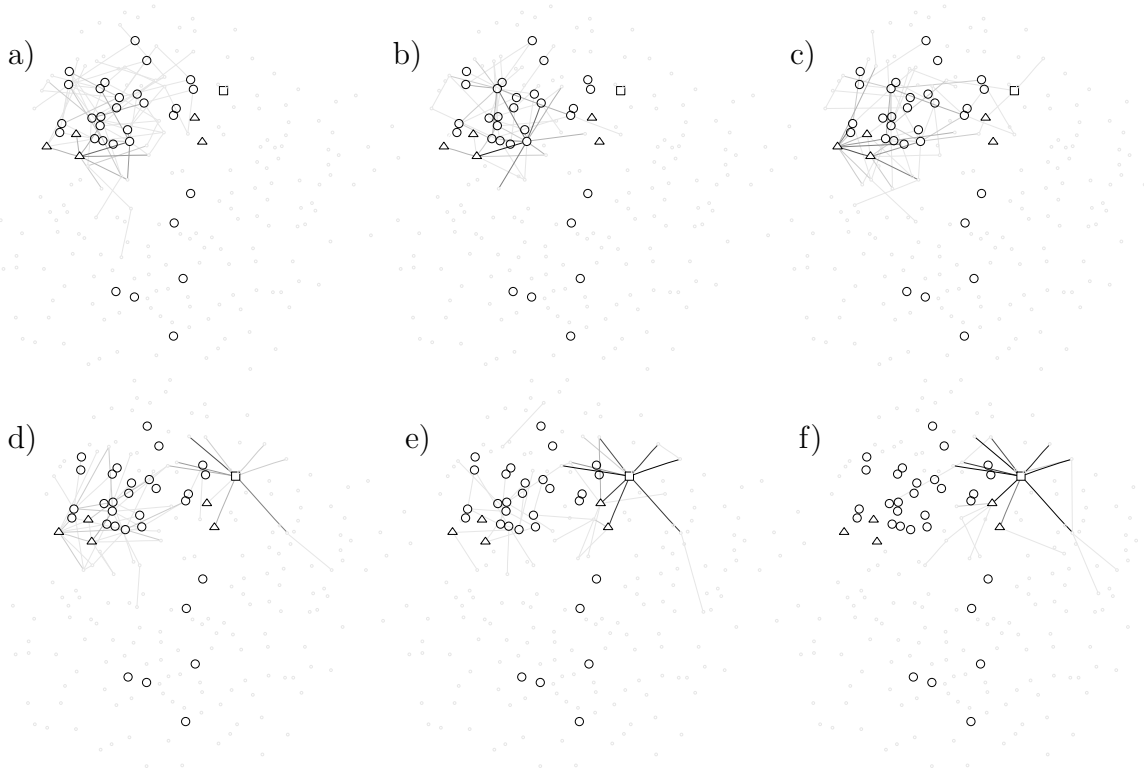


Figure 3.33: Emergence of the bifunctional network under periodically changing environmental conditions. In parts a–f the mutations over 6 consecutive periods of 100 generations are shown, starting with generation 701.

3.6.2 Interaction of networks by supply and demand of substrates

Imposing environmental conditions means that compounds produced by the networks do not change the environment. In the following this restriction is lifted and it will be taken into consideration that, firstly, the external resources are gradually consumed by the metabolic activities of the networks and, secondly, end products of a given network can be metabolised by other networks. In this way the environmental conditions for each network become time dependent which leads to a mutual dependency of the population and its environment.

As before, the process is modelled by an evolutionary algorithm as described in section 1.3. For the fitness function ϕ the special assumptions are made that $\phi = 1$ applies for a network which can metabolise a compound present in the environment and $\phi = 0$ for one that can not. Only network functions converting larger into smaller compounds are allowed for. Multifunctional networks are assumed to perform that

function for which the concentration gradient of the corresponding external compounds is maximal. The consumption rates are considered to be independent of the type of function.

The special case is considered where initially only C_6-compounds are present in the environment and it is assumed that the original population consists entirely of copies of one specially chosen network with the functions $\langle 6, 5\rangle$, $\langle 5, 4\rangle$ and $\langle 5, 2\rangle$. The population size is again set to $n = 100$ and the mutation rate per generation is set to $p = 0.01$.

Fig. 3.34 shows the time course for the amounts of all compounds available in the environment. As expected from the network function $\langle 6, 5\rangle$, there is an initial

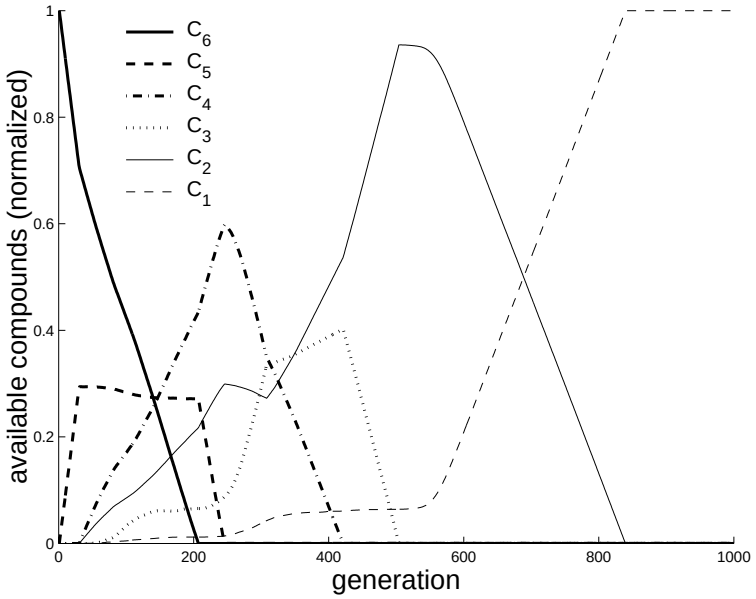


Figure 3.34: Time course of an evolutionary process taking into account the interaction of networks by supply and demand of substrates. Shown are the amounts of all compounds available in the environment. Initially, a limited amount of C_6 is present whereas all other compounds are produced in later stages of the process.

decrease in the amount of C_6-compounds and a simultaneous accumulation of the product C_5. Since the original network can metabolise C_5-compounds into C_4 and C_2, these compounds appear as soon as enough C_5 is produced (generation 33). The other compounds C_3 and C_1 cannot be produced by the original network and appear, therefore, only in later stages of the process after networks were discovered by mutation processes which can produce the corresponding compounds. The first compound which

is totally consumed is C_6 (after about 210 generations). The time courses for the amounts of C_6 and C_1 show a monotonous decrease and increase, respectively, whereas the amounts of the other compounds display a maximum. The larger the molecules the earlier they disappear from the environment. The process is finished when all C_6-compounds have been converted into C_1-compounds.

3.7 Discussion

The model presented in this chapter provides a structural analysis of networks consisting of a class of reactions which are typical for cellular metabolism. For simplification the huge number of reactions occurring in the living cell[1] is replaced by a limited set of generic reactions which only describe changes in the carbon skeletons of the metabolites. Specifically, included reactions are the transfer of a group of carbons from one metabolite to another and reactions splitting a compound into two smaller molecules or the reverse process. Apart from these characteristics, reactions occurring in living cells may show additional properties, such as phosphorylations or oxidation of compounds and many others. However, our simplified approach allows for a complete analysis of all possible networks consisting of subsets of these generic reactions. Whereas the assembly of alternative networks have been described before (Mavrovouniotis et al. 1990; Mittenthal et al. 1998; Mittenthal et al. 2001), the present study gives completely novel insights into network properties. The study is based on a precise definition of network functions and a newly introduced concept of multifunctional networks. Moreover, for the first time structural similarities between networks have been investigated by considering exchanges of a single reaction classified as transitions and mutations. This allows to quantify structural differences between networks in terms of a distance measure. In this way structure-function relationships of metabolic networks can be analysed accurately leading to a quantitative concept of robustness against changes in the network stoichiometry. The consideration of network mutations allows the implementation of evolutionary algorithms to search for networks possessing specified functions.

For carbon skeleton reaction networks involving compounds consisting of maximally six carbons the main results are the following:

1. For a given size $r \geq 3$ the networks consisting of all possible bi-bi- and bi-uni-reactions are complete in the sense that there exist networks for any possible

[1] see e. g. http://www.genome.ad.jp/kegg/

function. Networks of smaller size have only a limited number of functions. Restricting the set of generic reactions to bi-uni reactions the networks become complete for $r \geq 4$. Networks consisting exclusively of bi-bi-reactions are all non-functional.

2. A considerable fraction of the networks are multifunctional, although for networks of size $r \leq 3$ there are more monofunctional than multifunctional networks.

3. Any two networks of the same size $r \neq 2$ can be transformed into each other by a series of mutations, where each intermediate step results in an elementary network performing at least one function. For $r = 2$ there exist pairs of networks which can only be transformed into each other by accepting replacements resulting in non-functional networks.

4. Any mutation belongs to one of five mutation classes which are specified by their effect on the network function. Most mutations applied to networks of size $r \leq 4$ result in a complete change of function whereas for networks of size $r = 5$ the majority of mutations involve a partial loss or a gain of functions.

5. On average, the carbon skeleton networks show a rather high robustness since the majority of transitions from an elementary network result in another elementary network, meaning that the networks generally remain functional upon stoichiometric changes. The robustness increases with increasing network size. There exists exactly one network of size $r = 5$ which is completely robust in the sense that any possible exchange of a single reaction leads to another functional network.

6. Networks with the same function can be grouped into "islands". Within these islands every two networks can be transformed into each other by a series of mutations without losing this function. Networks of size $r = 5$ show the special property that for each function there exists only one large island. This means that it is always possible to transform two networks with the same function into each other by steps of mutations with every intermediate network also being able to perform this function.

7. Evolutionary optimisation under constant environmental conditions leads to the development of network populations toward clusters around islands of networks with a selective advantage. Networks belonging to large islands are favoured compared to networks of small islands. Under periodic environmental conditions

network populations evolve towards multifunctional networks which maintain their selective advantages under the various circumstances.

The concept of multifunctional networks as developed in this work is of relevance for understanding the design of real metabolic pathways. For example, cellular energy metabolism is multifunctional in the following respects. One and the same external source of substrate, e. g. glucose, can be metabolised into different end products such as pentoses required for DNA-synthesis, into amino acids with three or four carbons or into carbon dioxide for maximising the energy yield. On the other hand, a wide variety of substrates differing in their number of carbon atoms may serve as external resources, like sugars, amino acids and fatty acids. However, for a proper extension of the concept of multifunctional networks to cellular metabolism one has to consider a wider spectrum of generic reactions. Apart from changes in the number of carbon atoms, isomerisations, oxidation and reduction, adding and removal of phosphate groups and other functional groups can be included.

In the present study we make extensive use of graph theoretical descriptions. However, this description is not directly applied to the reaction networks themselves but to characterise the interrelations between networks regarding differences in their stoichiometry. The individual networks are mathematically described by the coefficients of their stoichiometric matrices. This allows to account in detail for the fluxes within these networks and to analyse their functions in terms of possible chemical conversions. In future work the gained knowledge will be helpful to extend the analysis to characterise the networks not only with respect to their stoichiometries but also to their kinetic behaviour. In this way a complete insight into different dynamic properties such as stability, oscillations or even chaos in carbon skeleton networks will be gained.

In this work we relate the metabolic function of a network to the conversions which can be performed in an elementary way. This is reasonable in view of the fact that nature will eliminate unnecessary reactions in the course of evolution. However, the results of our work show that networks which are elementary with respect to certain functions may perform other chemical conversions in a non-elementary way. In this case the elimination of single reactions which are unnecessary for non-elementary conversions would destroy other functions of the network. We assume that the mixed occurrence of elementary and non-elementary conversions is an important feature of real metabolic networks.

Chapter 4

Suggestions for future projects

The calculations presented in the previous chapters 2 and 3 have been facilitated by computer software developed as a part of this research project. The experience gained by the development of the software used for all simulations performed in chapter 2 led to a far more elaborated approach resulting in a very flexible, modular software architecture which has been used to produce the results presented in chapter 3. However, the system is more powerful and its capabilities are far from being exploited.

The software architecture is designed as a language extension of the programming language Perl strongly making use of its object oriented capabilities. It has intentionally been designed to be very flexible and further extensible. In this chapter several possible enhancements shall be suggested which can all be included in the software architecture with relatively little effort. In order to assess the possibilities and to get an impression on the scope of labour time needed, in some cases extensions have already been implemented. In the corresponding sections the approaches will be presented. However, these extensions are still in a developmental state.

In order to be able to describe the possible extensions of the computer model, the first section in this chapter will give an overview of the design of the software architecture and provide basic descriptions of the defined object classes. A detailed documentation of the packages is presently being written. This documentation is intended for other programmers who want to use and extent and / or modify the existing packages.

4.1 Software architecture

In the following a short description of the relevant object classes and their most important methods will be given. The intention is to give a short overview for the reader to support the basic understanding of the software system.

4.1.1 Biological object classes

An overview of the object classes representing objects with biological meaning is depicted in Fig. 4.1. In the nomenclature of Perl object classes the name of the class is

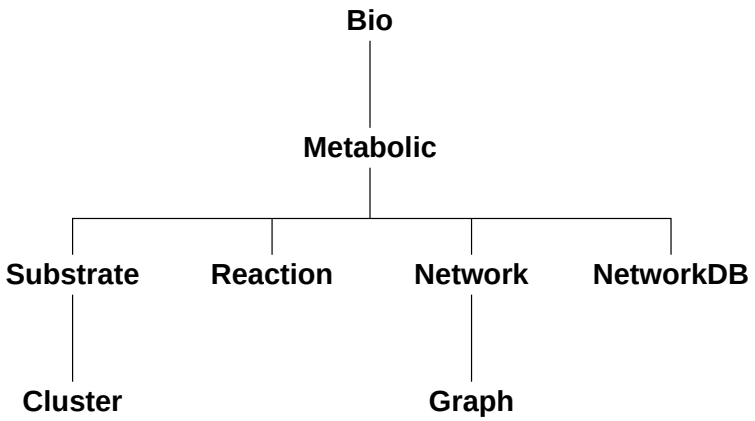


Figure 4.1: Object classes representing biological entities.

derived by joining the names of the subclasses as shown in the hierarchical tree from top to bottom with two colons.

Bio::Metabolic::Substrate

Instances of this class are used to describe a (bio-)chemical compound. It is defined by an arbitrary number of attributes. The methods `get_attribute` and `set_attribute` are the corresponding accessor methods. In the implementation used in chapter 3, the only attribute defined is the number of carbon atoms of the compound, denoted by **c**. Keeping the number of attributes completely free allows for a great flexibility and a wide range of possible implementations depending on the specific requirements of the model under investigation.

Bio::Metabolic::Substrate::Cluster

An instance of this class is defined by a list of instances of the class Bio::Metabolic::Substrate.

Bio::Metabolic::Reaction

An instance of this class defines a biochemical reaction. The reaction is specified by two instances of the class Bio::Metabolic::Substrate::Cluster which are accessed by the methods `in` and `out`. The method `kinetics` optionally provides access to a mathematical description of the reaction kinetics. The mathematical function is stored as a Symbolic::Function object, which is another language extension developed in the scope of this work – see section 4.1.2 below. So far, only multi-linear kinetics have been implemented and the only parameters that can be adjusted are equilibrium and rate constants. Possible extensions are provided for and it is planned to include the description of other kinetics (Hill, Michealis-Menten etc.).

Bio::Metabolic::Network

This is the central object class which describes a biochemical reaction network. It is simply defined by a list of instances of the class Bio::Metabolic::Reaction. It provides plenty of useful methods.

The distance between two networks is calculated by the method `dist`. In contrast to the definition given in section 3.4.1, this distance is simply defined as the number of reactions of the larger network minus the number of reactions which occur in both networks. In order to determine the distance defined in section 3.4.1, the representation of graphs as objects is required. For this purpose, the object class Graph::Undirected[1] has been used and extended by adding corresponding methods.

The stoichiometric matrix is available through the method `matrix`. It is returned as a `PDL`[2] object. The very powerful Perl Data Language (PDL) provides many matrix calculation methods amongst other useful tools. The module has been extended further by algebraic methods which were needed for the calculations performed to produce the results in chapter 3.

The method `can_convert` requires to substrates (instances of the object class Bio::Metabolic::Substrate) which are considered external and yields the solution vec-

[1] Author: Jarkko Hietaniemi, copyright 1999, O'Reilly & Associates, freely available at http://www.cpan.org

[2] Author: Christian Soeller, homepage: http://pdl.perl.org

tor (or matrix, if there is more than one linearly independent solution) of the steady state condition (3.2).

Two methods which are very useful to include dynamic behaviour in the analysis are `ODEs` which returns the set of differential equations (as a Symbolic::Function object) governing the reaction system based on the kinetics which are specified in the Bio::Metabolic::Reaction instances and `mfile` returning a text file which is understood by Matlab and can directly be used to numerically integrate the system equations.

Bio::Metabolic::Network::Graph

Instances of this class describe a graphical object visualising a network. The method `new_from_network` accepts one Bio::Metabolic::Network and two Bio::Metabolic::Substrate objects and generates a graphical representation of the network considering the two specified compounds as external. The generation of the graph is performed by the algorithm described in section 3.1.1.

The method `to_eps` returns an encapsulated postscript object which can be viewed by postscript viewers, printed or integrated in text documents.

Additional methods have been supplied to manually adjust the appearance of the graphical representation.

Bio::Metabolic::NetworkDB

This object class does not describe a biological object as such but rather provides access to a MySQL database. This allows the user to store networks together with any number of properties and access them any time later. However, since databases are not the focus of this work, this object class will not be further described.

4.1.2 Other object classes

Apart from the classes representing biological objects, some other useful classes have been developed in order to facilitate mathematical analysis.

Symbolic::Function

This is the most important class of the non-biological object classes. It is actually a front-end class for the subclasses described next. The tree representing the object classes is depicted in Fig. 4.2. It provides a symbolic representation of mathematical expressions. The subclasses are for internal use, representing constant values

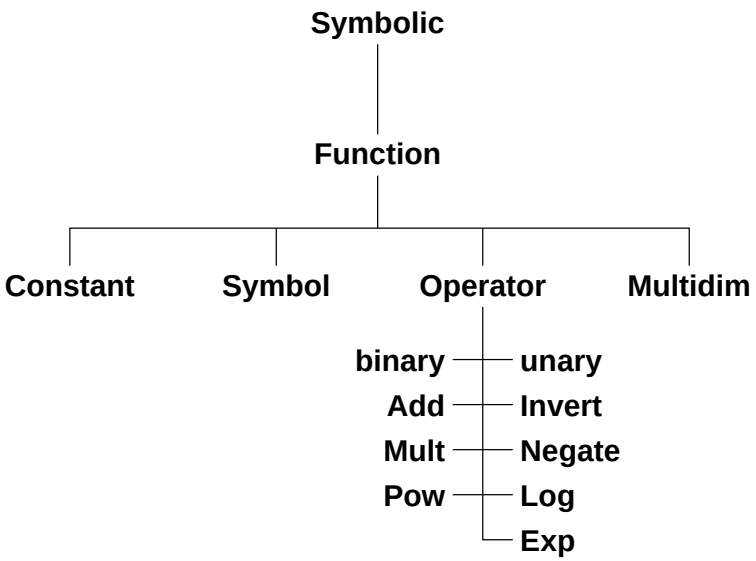


Figure 4.2: The object class and the subclasses used for representing mathematical expressions.

(Symbolic::Function::Constant), variables (Symbolic::Function::Symbols) and operators (Symbolic::Function::Operator and subclasses) such as addition, multiplication etc.

The subclass Symbolic::Function::Multidim represents a multi-valued function, e. g. a function $\mathbb{R}^n \to \mathbb{R}^m$.

The method `derive` expects a variable and returns another Symbolic::Function object representing the derivation of the function with respect to the given variable.

Based on this method, the method `newton` initiates a newton algorithm to search for the zeroes of the represented function. Starting values as well as precision limits can be provided. This method also works in the multi-dimensional case, even though the reliability is naturally limited.

The method `to_sub` returns a reference to a Perl subroutine which can be used for other calculations.

In order to evaluate a function with a given set of variable values, one uses the method `evaluate`.

It is also possible to simplify symbolic expression with the method `simplify` although this mechanism needs further elaboration.

Evo::Alg

This rather complex object class represents evolutionary algorithms as described in section 1.3. When creating an instance with the method `new`, several parameters have to be provided. First of all, the object class of the individuals is a prerequisite. Accordingly, a method for the random generation of individuals as well as a subroutine which performs the mutations on the individuals has to be provided. Optionally, a recombination procedure can be supplied. Other parameters that have to be specified are mutation probability, an evaluation subroutine, population size etc. In order to observe and control the process, maximally four callback subroutines can be specified which are called by the algorithm at four different stages within the main loop of the algorithm. Precisely, these four callbacks are initiated after each of the steps 2–5 of the algorithm described in section 1.3. These callback routines can also be used to influence the algorithm. For example, it is possible to alter external conditions from within these functions and thus indirectly influencing the evaluation of the individuals. By this means, the simulations for section 3.6.2 have been implemented.

After the successful creation of an Evo::Alg object, the only needed method is `step` which performs one loop (steps 2–5 in section 1.3) of the algorithm.

Matlab::Engine

This method is simply a wrapper around the Matlab C library to provide access to the powerful Matlab software package from within Perl subroutines.

The methods `new` and `Close` create and destroy the Matlab::Engine object through which the communication with Matlab is mediated.

The method `PutArray` transfers a multi-dimensional array from Perl to Matlab, the method `GetArray` retrieves a Matlab array and stores it in a Perl variable.

The only other method provided is `EvalString` which sends an arbitrary string expression to Matlab for evaluation.

These three methods are sufficient to make use of Matlab's capabilities from within Perl programs.

4.2 Further specification of the biochemical compounds

In chapter 3 a thorough analysis has been performed for metabolic networks consisting of reactions altering the number of carbon atoms of the participating compounds. For this purpose, the compounds have been defined only by their number of carbon atoms. In the computer representation this is reflected by assigning the instances of the object class Bio::Metabolic::Substrate one single attribute **c**. As mentioned in section 4.1.1, the choice of attributes is completely free. It is, for example, possible to describe the compounds by other attributes such as the number of phosphates bound to the molecule or the oxidation level of the compound.

Due to the combinatorial explosion, with these specifications a systematic approach as was followed in chapter 3 is impossible. This approach is rather interesting to combine the models and ideas from both chapters 2 and 3. By carefully selecting a set of generic reactions, e. g. phosphorylation / dephosphorylation with or without ATP, oxidation and reduction and changes to the carbon structure, far more realistic network structures might be produced by an evolutionary algorithm.

4.3 Models including the dynamic behaviour of networks

The models presented in chapter 3 only included static properties such as the stoichiometry of the reaction networks. This was very useful to classify networks with respect to their functions, i. e. their capability to interconvert certain external metabolites. However, it is evident that the dynamic behaviour of a network is also very important when investigating a network's biological properties. For example, two networks which can both perform a function $\langle a, b \rangle$ might in fact show a very different behaviour. Firstly, this concerns steady state conditions such as the flux through the system and the concentrations of the intermediate compounds which are both relevant quantities for cellular organisms. The steady state flux might indicate the efficiency of a biochemical reaction network to either produce substrates needed for other processes such as biosynthesis or to gain energy in form of ATP as was investigated in chapter 2. The concentrations of the intermediates should not become exceedingly large due to the limited osmotic pressure an organism or cell can suffer without taking damage. Secondly, the response to environmental changes is of interest. The dynamics

observed shortly after such changes indicates an organism's adaptability to different environmental conditions.

With the object classes and methods described in the previous section, it becomes clear that an extension of the investigation to the dynamical behaviour of reaction systems is not a very difficult procedure. The method `ODEs` of the class Bio::Metabolic::Network immediately yields the set of differential equations governing the reaction system. For example, using the method `newton` of the class Symbolic::Function, one immediately determines steady state conditions which in turn can be used to evaluate steady state fluxes.

For a first attempt in this direction, the following analysis has been performed: It is assumed that splitting carbon bonds yields (or costs) energy. As all bi-uni-reactions (reactions of type 2 – see section 3.1) split one carbon bond of a compound resulting in two smaller compounds, the equilibrium constants for these reactions are considered to assume the same value. Let q denote the equilibrium constant which for a reaction $C_j \rightleftharpoons C_k + C_l$ is defined as

$$q = \frac{[C_k][C_l]}{[C_j]}. \tag{4.1}$$

This quantity has the dimension of a concentration. All reference concentrations are considered to be one. Bi-bi-reactions (reactions of type 1) are considered to be energetically neutral, as they conserve the total number of carbon-carbon bonds. The corresponding dimensionless equilibrium constants are considered to be one.

In the following all networks of size $r = 3$ with the function $\langle 4, 1\rangle$ are examined. From Table 3.2 it can be seen that there are 31 such networks.

The concentrations of the external metabolites C_4 and C_1 are considered to be equal and are set to one.

It is now interesting to examine the steady state fluxes of the 31 networks for different values of q. Fig. 4.3 shows the steady state production rates of the external compound C_4 as a function of q for all considered networks. The values of q range from 10^{-3} to 10^3. As expected, the production rate is positive for all networks for values $q < 1$ (the building of larger compounds is energetically favourable) and negative for values $q > 1$ (the decomposition of compounds is energetically favourable) and there is no net flux for $q = 1$. The latter fact results from the choice of external metabolite concentrations. For both large and small equilibrium constants, the steady state flux generally seems to reach limits for the theoretical limits $q \to 0$ and $q \to \infty$. The only exception is one network for which the steady state production rate increases strongly with decreasing q. Closer inspection shows that this behaviour can be explained by the

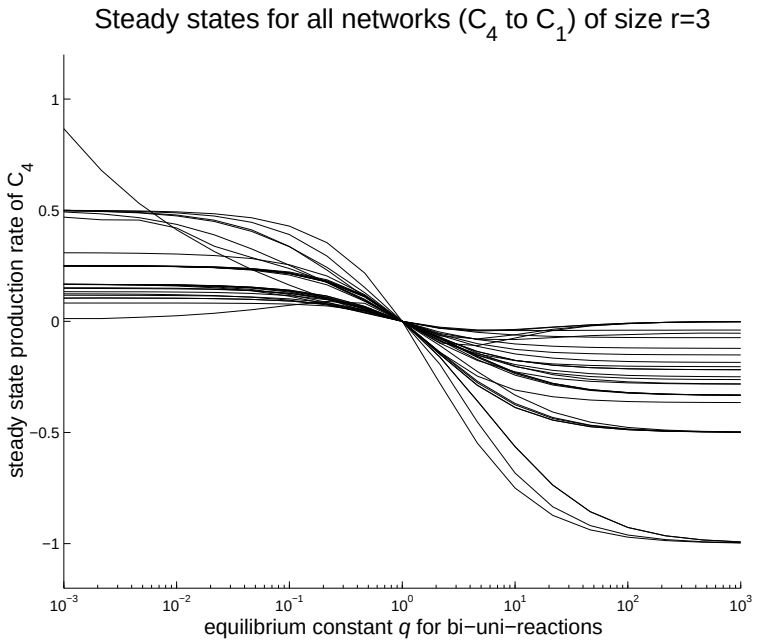


Figure 4.3: Steady state production (positive values) or consumption (negative values) rate for all networks of size $r = 3$ with the function $\langle 4, 1 \rangle$ as a function of the equilibrium constant q for bi-uni-reactions.

fact that the concentrations of the internal metabolites increase strongly. Since only a limited amount of substances can reside within a cellular organism, this behaviour is not realistic. Interestingly, the efficiency of the networks differ greatly. For low equilibrium constants (favouring the building of larger molecules), the limit for the production rate seems to be 0.5, whereas for large equilibrium constants (favouring the disintegration of large compounds into smaller molecules) there are four networks (the curves for two of these networks almost exactly coincide) which show a steady state consumption rate of close to 1. It is remarkable that the only network which shows a high performance for small as well as large equilibrium constants is the network with the simplest possible design for the task to interconvert the compounds C_4 and C_1. This network is shown in Fig. 4.4. This result indicates that simple designs might be favourable if one optimises for efficiency with respect to a high steady state production or consumption alone.

It becomes clear that with the capability of the developed software package a great variety of kinetic analyses can be performed. For larger sets (e. g. the graph G_5),

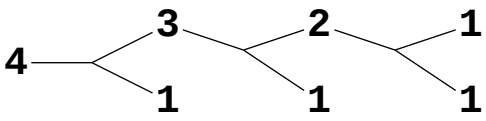


Figure 4.4: The most simple design for a network with $r = 3$ reactions and the function $\langle 4, 1 \rangle$. This network shows a good performance interconverting the compounds C_4 and C_1 regardless of the value of the equilibrium constant q for bi-uni-reactions.

evolutionary optimisation can be used to determine the most efficient network design for a specified task.

Making further use of the extension Matlab::Engine enables us also to investigate a time-dependent behaviour of the reaction systems. It is, for example, possible to observe the behaviour of multifunctional networks when switching from one external substrate to another. Using the results from Matlab, it is in principle possible to determine the efficiency of a network to adapt to new conditions. Of course, the hardest task is to define meaningful objective functions characterising this efficiency. One possibility is to measure the time the system needs to get close to the new steady state.

4.4 Regulatory mechanisms and enzyme activity

Since the kinetics of the reactions described by Bio::Metabolic::Reaction objects is implemented to be very flexible, it is also possible to include regulatory mechanisms in the models. For example, negative and positive feedback mechanisms can be introduced by making the reaction rates dependent on substrate concentrations. In the same way, varying enzyme activities can be simulated.

Similar to work on optimisation of kinetic activities and concentrations of enzymes by Heinrich et al. (1987), enzyme activity can easily be made the subject of an evolutionary algorithm by making use of the module Evo::Alg.

As a suggestion for first simulations in this direction, one suitable network design should be chosen and kept fixed. However, the enzymatic activities modelled by factors in the rate equations for the single reactions can be considered to be variable quantities and mutations can be defined which result in changes of the single enzyme activities. By this method, the present software architecture should be able to determine optimal

activities for a given network design. Again, the most crucial part is to define a biologically relevant objective function.

In a next step, both aspects could be combined and as well stoichiometric as enzymatic properties could become subject to an optimisation procedure.

4.5 Membranes and compartments

In the models described so far, biochemical reaction networks are considered "on their own", i. e. without any direct interaction with the environment. Environmental influence has been imposed by keeping concentrations of the external metabolites at a fixed level. In real cellular systems, however, the interaction with the environment takes place through membranes which keep certain compounds inside or outside the cell and permit the influx and outflux of others. First work in this direction has been already performed. A new object class Bio::Metabolic::Cell has been defined to simulate a simplified "cell" in an environment. The instances of this class are defined by a reaction network (Bio::Metabolic::Network), an environment which is essentially a list of substrates together with a concentration, and a list of permeabilities which define to what extent the exchange of the substrates is possible through the cellular membrane. To guarantee a high flexibility, the permeabilities are Symbolic::Function objects, meaning they are not necessarily defined by a fixed value but rather by an analytical expression which can be a function of all system variables such as the internal metabolite concentrations. The method `odes` returns the complete set of ordinary differential equations governing the behaviour of the network interacting with the environment through a membrane.

The ideas mentioned in section 4.4 have partly been realised by including control mechanisms in the object class Bio::Metabolic::Cell. The mechanisms controlling the flux through the membranes are simulated by linking the permeabilities to instances of the new object class Bio::Metabolic::Control which is essentially a simplified version of the class Symbolic::Function (see section 4.1.2). The regulatory rules define the permeabilities of the different compounds as functions of the intracellular metabolite concentrations. These functions are stored as tree-like structures for which mutation rules have been defined very similar to the methods used for genetic programming (see e. g. Banzhaf et al. 1998).

The following rules for the genetic programming approach have been chosen: All

internal metabolite concentrations may act as "input signals" of the form

$$f(x) = \frac{x^n}{K^n + x^n}, \tag{4.2}$$

where x denotes the metabolite concentration and K and n are two parameters defining the sigmoidal shaped signal $f(x)$. The parameter n corresponds to a Hill coefficient reflecting the steepness of the curve and K corresponds to the half-response value. All functions of the form (4.2) have in common that $f(0) = 0$, $f(K) = 1/2$, and $\lim_{x\to\infty} f(x) = 1$. The parameters n and K are subject to mutations. These signals can be combined to result in an "output signal" defining the permeability of the membrane for a specific compound. Possible combinations $g(x)$ of two signals $f_1(x)$ and $f_2(x)$ are given by

$$g(x) = f_1(x) \cdot f_2(x), \tag{4.3}$$

resembling a logical AND operator (the signal $g(x)$ is close to one if and only if both signals $f_1(x)$ and $f_2(x)$ are close to one) and

$$g(x) = 1 - (1 - f_1(x)) \cdot (1 - f_2(x)), \tag{4.4}$$

resembling a logical OR operator (the signal $g(x)$ is close to one if at least one of the signals $f_1(x)$ and $f_2(x)$ are close to one). Additionally a signal $f(x)$ can be negated (NOT operator) to result in the signal

$$g(x) = 1 - f(x). \tag{4.5}$$

These operators are combined to resemble a tree-like structure with the operators (AND, OR and NOT) as nodes and input signals as terminals. To allow for continuously opened or closed channels the terminals ON ($f(x) = 1$) and OFF ($f(x) = 0$) are also allowed. One such operator tree always defines a function accepting the internal metabolite concentrations as input values and yielding a real number between zero and one as return value. For these operator trees crossover rules can easily be defined by exchanging subtrees. This definition ensures that crossover operations always results in (mathematically) meaningful expressions.

As the focus of the present work lies on the complete analysis of different network designs, only some preliminary simulations have been performed to ensure that this approach is indeed promising.

The following scenario has been examined exemplarily: There exists one network of size $r = 3$ which can perform the function $\langle 6, 1\rangle$ as well as the function $\langle 4, 1\rangle$. This network played an important role in the simulations presented in section 3.6.1. It consists

of the reactions $(0, 2|1, 1)$, $(0, 6|2, 4)$, and $(2, 6|4, 4)$. We consider cellular organisms consisting of a membrane containing the three corresponding kinds of enzyme. The organism is supposed to "live" by metabolising either the compound C_6 or C_4 into C_1 while using the energy won in splitting the carbon bonds (cf. to section 3.6.1). We assume that the organisms can "feel" the internal metabolite concentrations (C_1, C_2, C_4, and C_6). Further, the cell possesses some regulatory mechanisms to open or close channels allowing the exchange of selected compounds between the interior of the cell and the environment. Clearly, there are maximally four such channels, one for each compound. In contrast to previous investigations where the structure of the metabolic system was considered to be subject to evolutionary processes, we here assume that the stoichiometry of the system remains constant but the regulatory mechanisms change due to selective pressure. The precise functions governing the regulatory mechanisms will be the output of the algorithm described above.

Additionally, the following assumptions have been made: Similar to the calculations presented in section 4.3, it has been assumed that reactions splitting a carbon chain into two shorter chains is energetically favourable. Specifically, for the reactions $(0, 2|1, 1)$ and $(0, 6|2, 4)$, the equilibrium constants defined analogously to Eq. 4.1 have been set to

$$q_{(0,2|1,1)} = q_{(0,6|2,4)} = 0.25. \tag{4.6}$$

This is a dimensionless number because concentrations are measured in reference units of one. This value means that the reaction $(0, 2|1, 1)$ on its own is in a state of equilibrium if the concentrations of the compound C_1 is 0.5 and the concentration of the compound C_2 equals 1. Similarly, the reaction $(0, 6|2, 4)$ is in equilibrium if the concentrations of the compounds C_2 and C_4 equal 0.5 and the concentration of the compound C_6 equals 1. The remaining reaction $(2, 6|4, 4)$ is considered to be energetically neutral. Assuming that the environmental conditions are such that the compound C_6 exists with a concentration of one and all other compounds are absent, clearly a favourable regulatory mechanism for an organism is to simply open the channel for the compounds C_6 (to let the energy source diffuse into the cell) and C_1 (to get rid of the waste product) and close the other two channels (to avoid any losses of intermediate substrates). With the given parameters, calculations reveal that under steady state conditions such an organism will split 0.655 carbon bonds per unit time. However, if environmental conditions change such that the resource C_6 is no longer available but rather the compound C_4 is present with the concentration of one, such an organism is doomed because it cannot make use of the new resource. An analogue organism which permanently opens the channel for C_4 (but not for C_6) will split 0.463 carbon bonds per unit time. However,

it is not able to make use of the compound C_6.

Therefore the following scenario has been investigated: Similar to the simulation in section 3.6.1 with changing environmental conditions, we also assume that periods alternate in which the compound C_6 (but not C_4) is available and vice versa. The periods have been set to 100 units of time. The fitness of an organism has been defined to be the integrated flux through the reactions splitting carbon compounds. This definition includes the efficiency of the organism to adapt to the two environmental conditions. The integration has been carried out over 400 units of time, meaning that the organism has been exposed twice to each environment. A simple regulatory design is to permanently open the three channels for C_6, C_4 and C_1. Then the organism can make use of both external resources but it must cope with a loss of intermediates. The performance value of this design amounts to 0.356 carbon bonds per unit time and can be used to compare the efficiency of different regulatory designs.

The evolutionary algorithm yields a regulatory structure which is considerably more efficient in making use of the alternating resource. The interesting fact is that all optimal structures seem to have in common that the design of the regulatory mechanism is very simple even though the algorithm allows for rather complex functions to be created. The optimal regulatory mechanisms of the cellular organism is visualised in Fig. 4.5. Interestingly, the only compound which has a regulatory effect is C_6. The

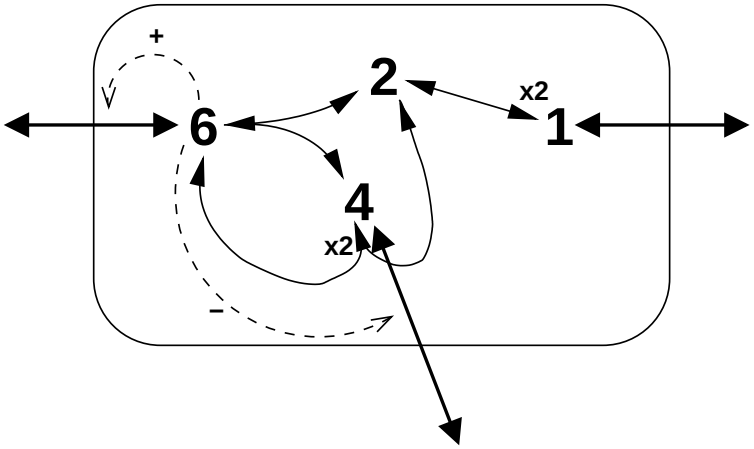


Figure 4.5: The optimised regulatory mechanism of a cellular organism under changing environmental conditions. The resources C_6 and C_4 are alternately available. The stoichiometry of the biochemical reaction network is predefined.

presence of this compound has a positive feedback effect on its own permeability p_6

through the membrane, whereas its presence negatively influences the permeability p_4 of the membrane for the compound C_4. As was expected, the channel for C_1 is permanently open whereas the channel for C_2 is permanently closed. The response curves for the permeabilities p_4 and p_6 have been determined by the algorithm and are given by the expressions

$$p_4 = 1 - \frac{[C_6]^{2.170}}{0.223^{2.170} + [C_6]^{2.170}}, \tag{4.7}$$

$$p_6 = \frac{[C_6]^{2.286}}{0.397^{2.286} + [C_6]^{2.286}}. \tag{4.8}$$

The response curves are plotted in Fig. 4.6. The performance of the organism with

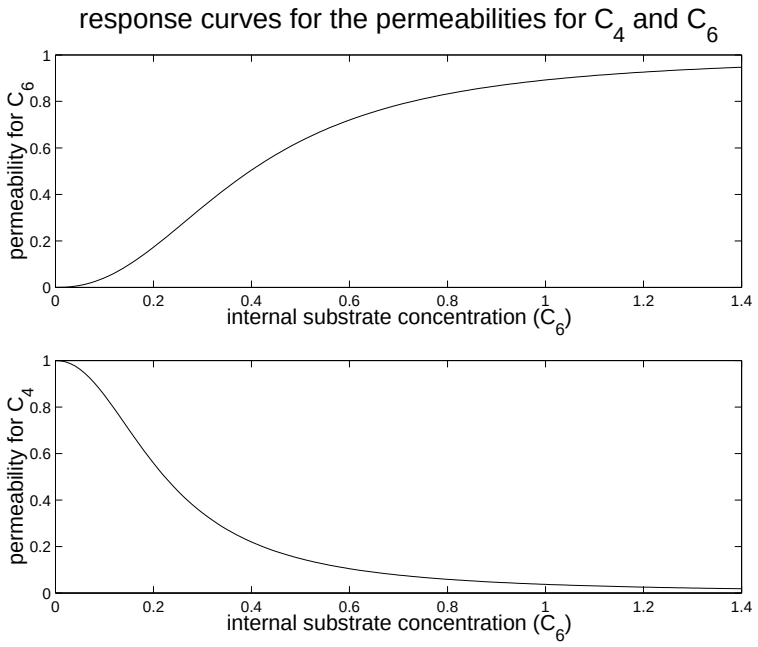


Figure 4.6: Response curves of the permeabilities for C_4 (bottom) and C_6 (top) as functions of the internal concentration of the metabolite C_6 for the optimised organism.

regulatory mechanisms governed by Eqs. 4.7 and 4.8 corresponds to an average of 0.445 carbon bonds split per unit time. This value is considerably larger than for the simple design with permanently opened channels.

Interestingly, if one removes the regulatory mechanism of the permeability of C_6 (but maintaining the mechanism for C_4) from the optimal design and replaces the

channel with one being permanently open, the performance only slightly decreases to 0.426 carbon bonds per unit time.

Table 4.1 summarises the important performance values for the five organisms mentioned. It is no surprise that the specialised organisms which can make use of only one

Simulated organism	availability of C_6	availability of C_4	alternating availability
Channel for C_6 open channel for C_4 closed	0.6546	0	0.3360
Channel for C_4 open channel for C_6 closed	0	0.4631	0.2334
Both channels for C_4 and C_6 open	0.4673	0.2406	0.3559
Channel for C_6 open channel for C_4 regulated	0.6346	0.2143	0.4260
Optimal design, both channels for C_4 and C_6 regulated	0.6183	0.2716	0.4454

Table 4.1: The splitting rates of carbon bonds for the simulated organisms. The rates are given for steady state conditions when either the compound C_6 or the compound C_4 is present as well as the average values for changing environmental conditions when these two compounds are alternately available.

of the resources C_4 and C_6 yield the highest steady state consumption rate if the corresponding resource is permanently available. For the case of the C_6-resource, the optimised organism yields a steady state rate not much lower than for the specialised organism. For the case of the C_4-resource, however, the steady state consumption rate is considerably lower than the for the specialised organism. The difference is understood by examining the steady state concentrations for the compound C_6. Computational analysis reveals that in case of the C_6-resource this concentration amounts to $[C_6]^* = 0.812$ resulting in a permeability for C_6 of $p_6([C_6]^*) = 0.837$ and for C_4 of $p_4([C_6]^*) = 0.057$ meaning that the channel for C_6 is almost completely open and the channel for C_4 is almost completely closed. In case of the C_4-resource the steady state concentration of the compound C_6 amounts to $[C_6]^\dagger = 0.237$ resulting in a permeability for C_6 of $p_6([C_6]^\dagger) = 0.235$ and for C_4 of $p_4([C_6]^\dagger) = 0.467$. This means that the channel for C_4 is not even halfway opened while still a considerable amount of the intermediate C_6 can leak through the corresponding channel. Is is interesting

that the algorithm did not discover more efficient parameters. Since more energy can be won from the C_6-resource per unit time, it is more important for an organism to improve the C_6-yield than the C_4-yield concerning its overall fitness under the given optimisation criteria. Therefore it can be assumed that the given parameters represent an optimal compromise for an organism which can make use of both substrates and therefore shows a high adaptability to changing environmental conditions.

It is remarkable that simple designs for regulatory mechanisms seem to be favourable even though in the simulations of the presented model no cost for a more complex design has been introduced. It can therefore be assumed that as a general tendency optimal regulatory structures can be realised by simple mechanisms.

However, to support this hypothesis further analysis is required. Since the continuation of investigations in this direction easily yields enough material of the scope of the present work, these first results at this place shall be understood as a suggestion for follow-up research projects. The calculations presented above give rise to optimism that this approach is indeed a step into the right direction to gain understanding of the structural design of regulatory mechanisms using optimisation principles.

It might be of further interest to simulate cellular systems not simply as interacting with the environment through a membrane but themselves consisting of several compartments. Based on the object class Bio::Metabolic::Cell, there is no principle difficulty to define a new object class which describes cellular systems consisting of several compartments which communicate through membranes with each other and through one membrane with the outer world. In this way one accounts for the compartmentisation of real organisms and simulations of cooperative metabolic systems occurring in different compartments like mitochondria and the cytosol should be possible.

4.6 Cell populations

In the previous section some possibilities to simulate cellular organisms interacting with their environment have been presented. It is of further biological relevance to simulate the behaviour of whole cell populations. Even this can be accomplished by extending the present software architecture. By using the flexible class Evo::Alg to simulate either an effect of the cells on their environment (like excretion of end products – c.f. section 3.6.2) or a direct influence (like predator and prey behaviour), it is only a small step from the object classes defined in the previous section to simulations of this kind.

4.7 Signal transduction pathways

Recently, the theoretical study of signal transduction pathways gained increasing interest. First model approaches were concerned with the description of ultra-sensitivity in protein kinase cascades (Huang and Ferrel 1996). Lately, the structural analysis of these pathways became subject of research where especially questions on the principle functioning of the differential feedback regulation of the MAPK cascade were raised (Brightman and Fell 2000).

From a mathematical point of view, these pathways are structurally similar to metabolic systems. Of course, there are elementary differences, first of all the fact that a signal transduction pathway transfers information (e. g. by a cascade of phosphorylations and dephosphorylations like in the MAP-kinase pathway) rather than performs chemical transformations as a metabolic reaction network does. However, the behaviour is governed by a set of ordinary differential equations depending on concentrations of metabolic compounds or ratios of such concentrations. Therefore there is no reason why these important biological mechanisms could not be studied using the present software architecture. In order to do so, corresponding object classes have to be defined to reflect signal transduction pathways. The regulatory mechanisms should be handled similarly to those described in section 4.5. A suggestion for first work in this direction is to chose a well known signal transduction pathway and then try to define a reasonable objective function, e. g. reflecting the stability of the signal. Then the optimal regulatory mechanism of the pathway could be determined using an evolutionary algorithm. In a next step, the topology of different signal transduction pathways can be compared regarding their efficiency and if one succeeds in finding reasonable mutation rules to change the structural design of such pathways, even structural optimisation should be possible.

Chapter 5

Conclusions

In this work two models have been presented both following the same underlying perspective. Both models have in common that they are directed to examine the structural properties of metabolic systems using optimisation principles. In chapter 2 the central energy metabolism of aerobic organisms comprising glycolysis and the citric acid cycle together with oxidative phosphorylation has been selected as the subject of research. A model has been specially designed for the purpose of representing alternative reaction sequences consisting of the same class of *generic* reactions as the real pathway. By defining mutation rules, an evolutionary algorithm has successfully been applied to determine the optimal structure of ATP and NADH producing pathways regarding their overall steady state ATP production rate. The optimised reaction sequences resemble the existing pathways in many ways which leads to the conclusion that a high ATP production rate was an important feature giving a selective advantage during the early stages of the evolution of cellular metabolism.

In contrast to the model presented chapter 2 which was limited to the description of unbranched reaction sequences, the model described in chapter 3 allows for the representation of branched network structures. Rather than specifying the model description for the purpose of investigating one selected metabolic subsystem, the focus of this model lies on a complete and thorough analysis of possible structural designs. The generic reactions have been chosen to represent enzymatic reactions which either split or merge carbon containing compounds or transfer a group of carbons from one compound to another. By particularising biochemical compounds by their number of carbon atoms alone, any metabolic system can be characterised by an underlying skeleton network consisting of the defined generic reactions. Precise definitions for such network's functions, stoichiometric similarities and their robustness with respect to sto-

ichiometric changes have been introduced. Based on these definitions it was possible to investigate a number of evolutionary scenarios. The simulations presented in this work include the development of a population of networks under constant as well as changing environmental conditions and the dynamics within a population of networks interacting through the supply and demand of metabolic compounds. A particularly interesting result is the reproduction of the development of multifunctional networks which have a selective advantage due to their adaptability to different environmental conditions. Concluding, it can be claimed that the presented model contributes to the comprehension of possible network structures regarding their stoichiometric properties as well as their biological functions. The gained knowledge is intended to serve as a basis for future research in the field of structural optimisation. In chapter 4 several possible extensions have been suggested and it becomes clear that the applicability of the developed model is not limited to the stoichiometric properties of network structures. Rather, first simulations have successfully been performed which consider the dynamic behaviour of the network structures (see section 4.3). Combining this approach with the two approaches of the models described in chapters 2 and 3, the structural analysis using optimisation principles can be extended to a vast number of metabolic systems like the pentose phosphate pathway or the Calvin cycle which cannot easily be simplified to be represented by an unbranched structure. The most challenging task when pursuing investigations in this direction will be to find suitable objective functions for the optimisation process or defining realistic scenarios. However, the first results that have been presented suggest that this is a very promising approach.

Apart from stoichiometric properties, the behaviour of metabolic systems is decisively determined by the regulatory mechanisms controlling the activities of the single enzymes as well as transport mechanisms through cellular membranes. In section 4.5 first results have been presented regarding the structural analysis of regulatory mechanisms. It has been shown that regulatory mechanisms can also be subject to optimisation principles and the results give rise to optimism that the structural designs of these mechanisms can be understood by applying the same underlying principles that the models described in chapters 2 and 3 are based on. A possible route to success regarding this field of research has been outlined.

Summarising, it has been confirmed that optimisation principles show a wide applicability in the study of biochemical and biophysical systems. The insight gained concerning the fundamental structures of cellular metabolism may serve as a foundation for future studies in a great variety of research areas wherever the understanding of structures is a major concern. In the scope of the present work an extremely flexible

and powerful tool has been developed that certainly will facilitate ongoing research projects regarding metabolic optimisation and related fields.

Appendix A

Additional topics to chapter 2

A.1 Parameter choices

The purpose of this model is to reproduce and explain some fundamental structural properties of ATP producing pathways with as little *a priori* input as possible. Therefore we keep the choices of the parameters rather unspecific.

Concerning the values of the thermodynamic equilibrium constants, we assume that in a reaction chain C, ATP consumption, NADH consumption and dephosphorylations are thermodynamically favourable reactions, i. e. q_A, q_N, $q_p > 1$. Furthermore, $q_U > 1$ should hold true, since the overall ATP production is driven by a drop in free energy from the initial substrate X_0 to the final product X_{r_C}. Protonisation and deprotonisation are considered fully reversible ($q_H = q_h = 1$). The degree of the reversibility of the other reactions is controlled by a parameter $\lambda > 0$, such that

$$q_A = {q_{A,0}}^{\lambda} \tag{A.1}$$

and similar expressions for all other equilibrium constants. Further, $q_a = 1/q_A$, $q_n = 1/q_N$, $q_P = 1/q_p$. All calculations were performed with $q_{A,0} = q_{N,0} = q_{p,0} = q_{u,0} = 10^3$ and with different values of λ. Note, that a change in λ means a linear rescaling of the free energy changes of the corresponding reactions.

We assume all occurring reactions – apart from the faster protonisation / deprotonisation – to process with the same characteristic time. We choose $\tau_u = 1$, $\widetilde{\tau_A} = \widetilde{\tau_a} = 1$, $\widetilde{\tau_N} = \widetilde{\tau_n} = 1$, $\tau_P = \tau_p = 1$ and $\tau_H = \tau_h = 0.01$. Measuring the time in units of hours, this choice is close to characteristic times of some glycolytic reactions, which are typically in the range of minutes up to an hour (Rapoport et al., 1976; Joshi and Palsson, 1989, 1990). The characteristic times $\widetilde{\tau_x}$ in the reference states enter

expressions such as

$$\tau_N = \widetilde{\tau_N} \cdot \frac{1 + q_N}{2\left((1 - n_2) + q_N \cdot n_2\right)} \geq \widetilde{\tau_N}/2 \tag{A.2}$$

for NADH consuming reactions. The analogue is valid for τ_A – see Eq. (2.11) –, and for τ_a and τ_n.

Moreover, we assume for reasons of simplicity, $A = N = 1$ and $X_0 = 1$, which approximately corresponds to experimental values if concentrations are measured in mM units.

An appropriate choice for the value of k_d is obtained by the following estimation: Assume, J_d is supposed to be such that it can cope with any excess production of NADH through J_C. The forward fluxes $J^+_{C,i}$ are given by

$$J^+_{C,i} = \frac{X_{i-1} \cdot q_i}{\tau_i(1 + q_i)} \leq \frac{X_{i-1}}{\tau_i}, \quad \text{for each} \quad i = 1 \ldots r_C. \tag{A.3}$$

Thus,

$$J_C \leq J^+_{C,1} = \frac{X_0 \cdot q_1}{\tau_1(1 + q_1)} \leq \frac{X_0}{\tau_1}. \tag{A.4}$$

We first exclude the case that the first reaction is of type 'h'. In case the first considered reaction is of type 'A', 'a', 'N' or 'n', the inequality $\tau_1 \geq 1/2$ holds true because of Eq. (A.2) or its equivalent – depending on the type of the reaction. For the other reactions ('P', 'p' and 'u'), $\tau_1 = 1 \geq 1/2$ also holds. Therefore, as $X_0 = 1$, the upper limit of J_C is given by

$$J_C \leq 2. \tag{A.5}$$

This upper limit is in good agreement to the glycolytic flux, for example in erythrocytes it is about 1 mM/h (Mulquiney and Kuchel, 1999a).

Now consider the case that the first reaction is of type 'h'. As this reaction is very fast (compared to all other reaction types), it can be considered to be close to equilibrium. Since $q_h = 1$, one gets $X_1 \approx X_0$ and the above estimation holds true by increasing all indices by 1. In case the first two reactions are of type 'h' (this is the maximum number of 'h'-reactions possible at the beginning of the reaction chain), the indices have to be increased by two and the same upper limit for J_C results.

In case of excess production of NADH we assume that approximately $n_2 = 1$, yielding $J_d \approx k_d$. Furthermore we assume that the reasonable limitation $n \leq 5$ holds, which has been confirmed by the simulations which yield $n = 4$ for the most efficient sequences (see section 2.3). Since $J_d \leq nJ_C$ – see Eq. (2.16) –, we fulfil our initial

assumption by the choice $k_d = 10$. As the NADH production resulting from J_C can be compensated by J_{Ox} as well as J_d, an appropriate choice for k_{Ox} is $k_{Ox} = k_d$.

From Eq.(2.22) we see that the parameter k_{ATPase} affects the behaviour of the dependency between a_3 and n_2. As expected, increasing values of k_{ATPase} will decrease the concentration of ATP. In our calculations we used $k_{ATPase} = 2$, which yields for the optimal sequences an ATP concentration neither too low nor too high compared with the total concentration of adenine nucleotides. In reality this fraction is $A_3/A = a_3 \approx 0.75$. The efficient sequences listed in Table 2.2 yield $a_3 \approx 0.9$. Moreover, the value of k_{ATPase} is close to the value for the rate constants of external ATP consuming processes in erythrocytes (Rapoport et al., 1976; Mulquiney and Kuchel, 1999a).

A.2 Number of elements in the space of reaction sequences

We numerate the ligand states by indices $1 \ldots 9$ as depicted in Fig. 2.1. We define the symmetric (9×9) matrix $\mathcal{M}$ by

$$\mathcal{M}_{ij} = \text{number of possible sub-pathways beginning in ligand state } i \text{ and ending in ligand state } j \text{ without changing the internal state, i. e. with no 'u'-reaction} \tag{A.6}$$

Having defined the matrix $\mathcal{M}$, we can calculate in the next step the number of possible pathways beginning in ligand state i, containing one 'u'-reaction and ending in ligand state k. We denote this number by $\mathcal{N}^{(1)}_{ik}$. A 'u'-reaction can have as substrate any of the ligand states, so the number of possible pathways is

$$\mathcal{N}^{(1)}_{ik} = \sum_{j=1}^{9} \mathcal{M}_{ij}\mathcal{M}_{jk} = \left(\mathcal{M}^2\right)_{ik}. \tag{A.7}$$

This expression can be generalised as

$$\mathcal{N}^{(U)}_{ik} = \left(\mathcal{M}^{U+1}\right)_{ik}. \tag{A.8}$$

Hence, we arrive at the formula for the number $N^{(U)}$ of possible pathways beginning and ending in ligand state 1 (equivalent to `HH` in section 2.1.1) and containing U 'u'-reactions

$$N^{(U)} = \left(\mathcal{N}^{(U)}\right)_{1,1} = \left(\mathcal{M}^{U+1}\right)_{1,1}. \tag{A.9}$$

We demonstrate the calculation of the matrix $\mathcal{M}$ by counting the possible sub-pathways from ligand state 1 to ligand state 2.

There are exactly eight topologically different routes from state 1 to state 2 (see Fig. A.1). Every single arrow can represent exactly two generic reactions (see the

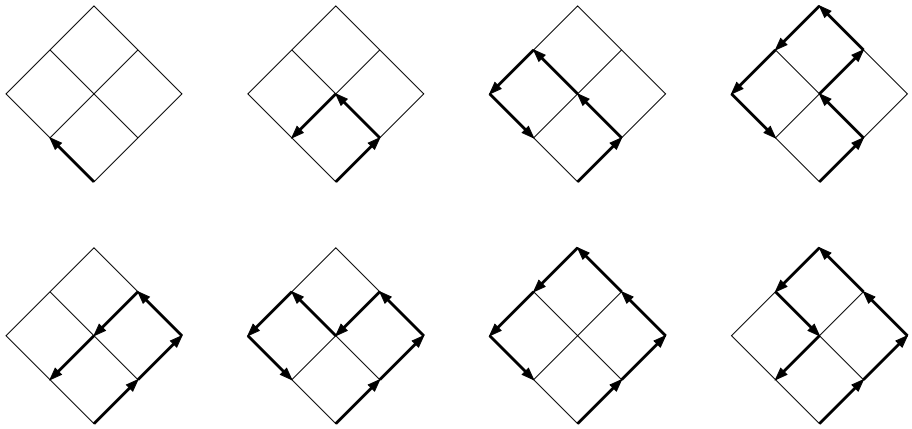


Figure A.1: Possible sub-pathways from ligand state 1 to ligand state 2 without 'u'-reactions

discussion on the topology of pathways in Appendix A.3), thus the number of pathways every graph in Fig. A.1 represents is 2^a, with a denoting the number of arrows. Thus,

$$\mathcal{M}_{1,2} = 2^1 + 2^3 + 2^5 + 2^7 + 2^5 + 2^7 + 2^7 + 2^7 = 586. \tag{A.10}$$

In the same manner we can calculate all other entries of $\mathcal{M}$. By definition, if the final and initial states are the same, there is only one possible pathway (the "empty" pathway), therefore,

$$\mathcal{M}_{ii} = 1. \tag{A.11}$$

We get

$$\mathcal{M} = \begin{pmatrix} 1 & 586 & 884 & 586 & 680 & 440 & 884 & 440 & 864 \\ 586 & 1 & 586 & 296 & 338 & 296 & 440 & 228 & 440 \\ 884 & 586 & 1 & 440 & 680 & 586 & 864 & 440 & 884 \\ 586 & 296 & 440 & 1 & 338 & 228 & 586 & 296 & 440 \\ 680 & 338 & 680 & 338 & 1 & 338 & 680 & 338 & 680 \\ 440 & 296 & 586 & 228 & 338 & 1 & 440 & 296 & 586 \\ 884 & 440 & 864 & 586 & 680 & 440 & 1 & 586 & 884 \\ 440 & 228 & 440 & 296 & 338 & 296 & 586 & 1 & 586 \\ 864 & 440 & 884 & 440 & 680 & 586 & 884 & 586 & 1 \end{pmatrix}. \tag{A.12}$$

In Table A.1, the number of possible pathways $N^{(U)}$ dependent on U are shown.

U (Number of 'u'-reactions)	$N^{(U)}$ (Number of possible pathways)
0	1
1	$3.8458 \cdot 10^6$
2	$1.3803 \cdot 10^{10}$
3	$6.4908 \cdot 10^{13}$
4	$2.8792 \cdot 10^{17}$
5	$1.2921 \cdot 10^{21}$
6	$5.7852 \cdot 10^{24}$
7	$2.5914 \cdot 10^{28}$
8	$1.1607 \cdot 10^{32}$
9	$5.1986 \cdot 10^{35}$

Table A.1: Number of possible pathways for a given number of 'u'-reactions

For large U, the approximation

$$N^{(U)} \approx \Lambda^{(U+1)} \tag{A.13}$$

holds, where Λ is the largest eigenvalue of $\mathcal{M}$

$$\Lambda = 4479.0 \tag{A.14}$$

The relation

$$\frac{N^{(U+1)}}{N^{(U)}} = \Lambda \tag{A.15}$$

is approximately fulfilled even for rather small values of U

$$\frac{N^{(6)}}{N^{(5)}} = 4477.2, \tag{A.16}$$

$$\frac{N^{(10)}}{N^{(9)}} = 4479.0. \tag{A.17}$$

A.3 Definition of the mutations and construction of alternative pathways

A given sequence is defined as an unbranched chain of reactions of type u, H, h, N, n, P, p, A and a. During the genetic algorithm we create new sequences from already existing ones by applying appropriate changes of the specific type and the order of the reactions. In the following we call these changes *mutations*. "Point mutations" are defined as a simple exchange of a single reaction by another one at a fixed location. Whereas exchanges of reactions within the pairs (H,N), (h,n), (P,A), (p,a) are always possible, other point mutations are not allowed. For example, a reaction cannot be replaced by 'A' if the target molecule is in an unphosphorylated state. Accordingly, not every possible reaction sequence can be created by using only point mutations. We define a minimal set of mutations that allows us to construct all possible reaction sequences. It contains point mutations as well as more complex alterations of the reaction sequence.

Let us define the *topology* of a sequence by a graph connecting the nodes of the reaction pathway, see Fig. 2.2 as an example. Such an arrow may connect either two molecules with the same internal state but different ligand states (by one of the coupling reactions) or two molecules with the same ligand state but different internal states. In the former case the nature of a given coupling reaction is not specified for a given topology. The topology only determines which above mentioned pair the reaction belongs to. According to our assumptions the latter case always denotes a 'u'-reaction.

In order to create a given reaction sequence with a fixed number of 'u'-reactions, we first construct a pathway having the correct topology using *topological* mutations. Afterwards, we specify the nature of the coupling reactions by applying point mutations.

We define the following classes of topological mutations (see Table A.2):

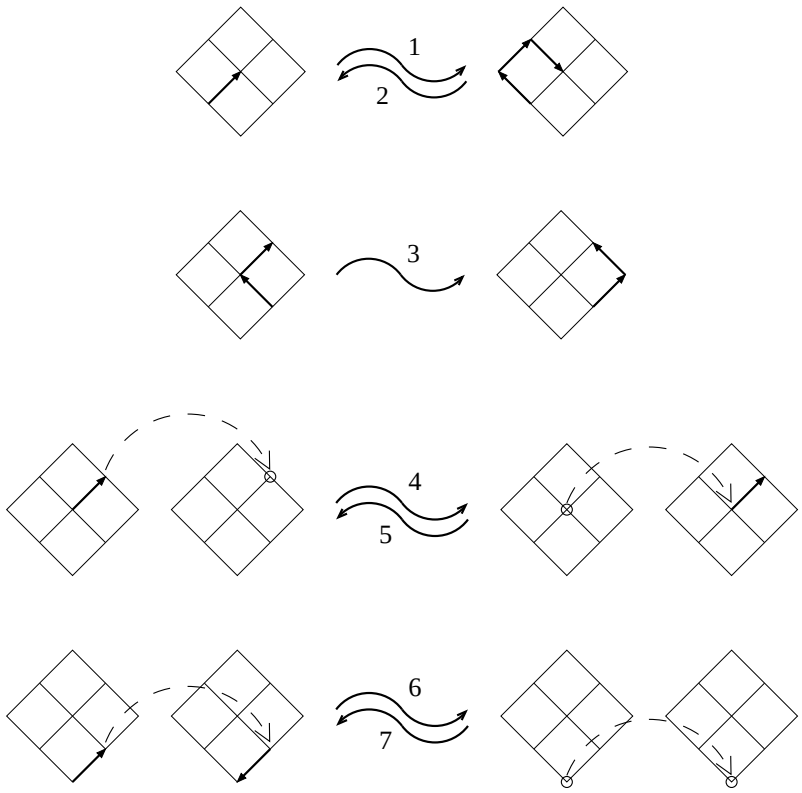


Table A.2: Schematic visualisation of mutation classes. Coupling reactions are denoted by straight arrows, the dashed arrow represents a ‘u’-reaction. Each S-shaped arrow denotes the effect of the corresponding mutation class. Classes 1–3 do not involve ‘u’-reactions, classes 4–7 do.

1. Replace one arrow by three, such that the first and third arrow point in the opposite direction and the second one points in the same as direction as the original arrow.
2. Reverse of 1.
3. Replace two arrows by two, exchanging their direction.
4. Exchange an arrow with the following ‘u’-reaction.
5. Reverse of 4.
6. Replace three arrows, the middle one of which represents a ‘u’-reaction by one ‘u’-reaction.

7. Reverse of 6.

The length of a reaction sequence is not changed by mutations 3, 4 and 5. The other mutations do change this length.

With these classes of topological mutations and the point mutations we can formulate the following

Theorem 1 *Any sequence I with U 'u'-reactions can be transformed into any sequence J with the same number of 'u'-reactions by a finite number of mutations of the classes 1–7 and point mutations.*

We prove this statement by constructing an arbitrary given pathway I from the sequence $\mathbf{uu\ldots u}$ by applying mutations. As any mutation is reversible, we then can transform any given sequence into the sequence $\mathbf{uu\ldots u}$ and transform this one into any other given sequence and our assumption is proven.

Because the mutations can be categorised into topological mutations and non-topological mutations (point mutations), we construct a pathway with the given topology using mutations of classes 1–7 and considering only the "topological" properties of a reaction, thus e. g. A and P reactions are equivalent. After such a pathway was constructed, it is obvious, how it can be transformed into the correct pathway by applying point mutations.

The construction of a pathway with a given topology takes place from the beginning to the end. Step by step the correct 'sub'-pathways between two 'u'-reactions are assembled. The mechanism is graphically illustrated in Table A.3 using a special case as an example.

In the following instruction how to construct a given pathway, we denote by I any pathway with the same topology as the given pathway. The 'sub'-pathways between the i-th and $i+1$-st 'u'-reaction are denoted by I_i. Consequently, I_0 denotes the subsequence before the first 'u'-reaction. Thus

$$I = I_0\mathbf{u}I_1\mathbf{u}\ldots\mathbf{u}I_U. \tag{A.18}$$

Note, that the maximum length of any I_i is 8 reactions. This restriction holds because no ligand state (there are 9 of them) can be passed through more than once without changing the internal state of the molecule, i. e. applying a 'u'-reaction.

The pathway created in the j-th intermediate step is denoted by $M^{(j)}$, with the starting sequence

$$M^{(0)} = \mathbf{uuu}\ldots\mathbf{u} \tag{A.19}$$

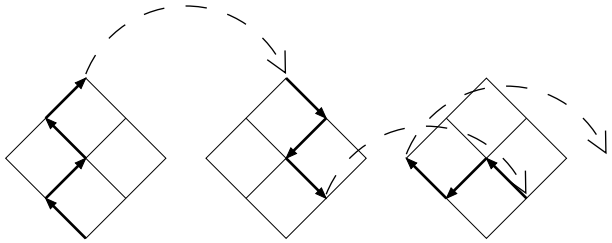
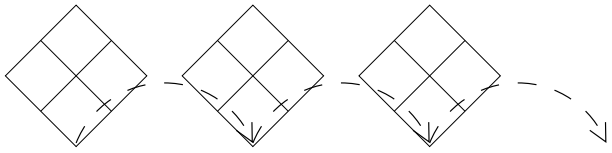
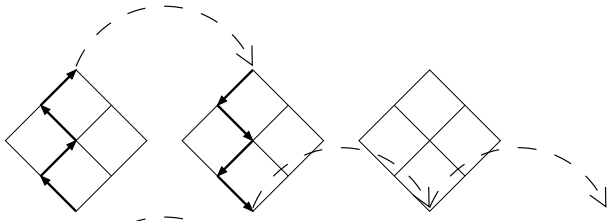
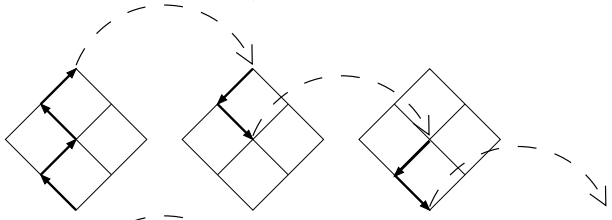
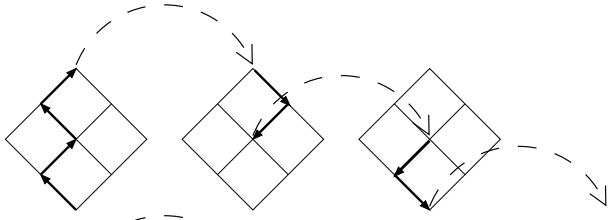
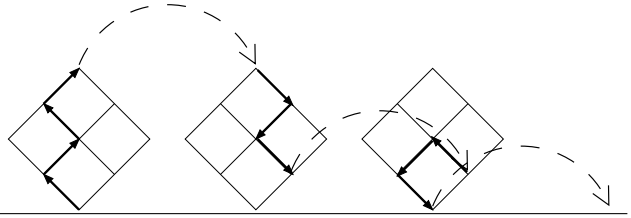
Topology of the given pathway	
Initial Pathway	
Step 1	
Step 2	
Step 3	
Step 4	

Table A.3: Construction of an arbitrary pathway. In four steps the topology of the first two sub-pathways are constructed. The next sub-pathways are subsequently constructed repeating steps 2–4.

Furthermore, let $\overline{K}$ denote the sequence consisting of the reverse reactions of K occurring in the opposite order.

The construction of I by applying mutations on $M^{(0)}$ takes place in several steps:

1. We first apply zero or more mutations of class 7 on $M^{(0)}$ in order to generate the sequence
$$M^{(1)} = I_0\mathbf{u}K_1\mathbf{u}\ldots\mathbf{u}, \tag{A.20}$$
with $K_1 = \overline{I_0}$.

2. Split both pathways I_1 and K_1 into two sub-pathways $I_1'I_1''$ and $K_1'K_1''$ respectively, such that I_1' and K_1' end in the same ligand state. This is always possible, because I_1 and K_1 begin in the same ligand state, thus $I_1' = K_1' = \emptyset$ is always a possible solution. In case there is more than one possibility to split the pathways, choose the one which yields the longest I_1'. This instruction yields a unique division of I_1 and K_1. Now apply zero or more mutations of class 4 to generate the sequence
$$M^{(2)} = I_0\mathbf{u}K_1'\mathbf{u}K_1''\mathbf{u}\ldots\mathbf{u}. \tag{A.21}$$

3. The two sequences K_1' and I_1' have the same initial and final ligand states. Therefore it is possible to transform K_1' into I_1' by applying a necessary number of mutations of classes 1, 2 and 3 as can be easily seen from the definition of these mutation classes. This yields
$$M^{(3)} = I_0\mathbf{u}I_1'\mathbf{u}K_1''\mathbf{u}\ldots\mathbf{u}. \tag{A.22}$$

4. By applying mutations of class 7, generate the sequence
$$\begin{aligned} M^{(4)} &= I_0\mathbf{u}I_1'I_1''\mathbf{u}\overline{I_1''}\,K_1''\mathbf{u}\ldots\mathbf{u} && \text{(A.23)}\\ &= I_0\mathbf{u}I_1\mathbf{u}\overline{I_1''}\,K_1''\mathbf{u}\ldots\mathbf{u}. && \text{(A.24)}\end{aligned}$$
This is always possible because the subsequences K_1'' and I_1'' have been constructed in such a way (see step 2) that they both begin in the same ligand state but do not pass through any other common state. Thus, $\overline{I_1''}$ ends in the ligand state K_1'' starts in. We define
$$K_2 = \overline{I_1''}\,K_1'' \tag{A.25}$$
and write
$$M^{(4)} = I_0\mathbf{u}I_1\mathbf{u}K_2\mathbf{u}\ldots\mathbf{u}, \tag{A.26}$$

which is an analogue of Eq. (A.20) with one more subsequence (I_1) already correct.

5. Now repeat steps 2–4 to construct the subsequences I_i with $i = 2 \dots U$. The construction is obviously possible for $i = 2 \dots U-1$. For $i = U$ no mutations of classes 4–7 can be used, because I_U is not followed by a 'u'-reaction. Note, however, that in the last step ($i = U$), the final ligand state is the ground state, therefore the splitting of the sequences will yield $I''_U = K''_U = \emptyset$. Therefore, no mutations are needed in steps 2 and 4, i. e. no mutations involving 'u'-reactions are necessary.

After U iterations of steps 2–4 we get

$$M^{(3U+1)} = I_0 \mathbf{u} I_1 \mathbf{u} \dots \mathbf{u} I_U = I. \tag{A.27}$$

This sequence is of the same topology as the given pathway. Therefore, we can create the exact given pathway by specifying the reaction types, i. e. by applying point mutations and the theorem is proven.

A.4 Analytical tools

A.4.1 Distance between two sequences

Let us denote the sequence space by $\mathcal{S}$. In principle, any function

$$D \; : \; \mathcal{S} \times \mathcal{S} \mapsto [0, \infty) \tag{A.28}$$

that fulfils the conditions (2.32)–(2.34) can be used as a distance measure. Our goal was a simple definition that intuitively fulfils the aspects of a *distance*, i. e. "similar" sequences should yield a smaller distance than very "different" sequences.

The distance measure we define is in principle a refined Hamming distance (Hamming, 1980). First we define how any sequence string is rewritten such that all rewritten strings have the same fixed length. We number the possible coupling reactions from Fig. 2.1 from 1–12, for example as depicted in Fig. 2.1. Note, however, that the exact way of numbering the reactions is irrelevant. A given pathway I is divided into sub-pathways I_i as in Eq. (A.18) in Appendix A.3. Now, for every subsequence I_i a string J_i, containing exactly 12 characters, is constructed by placing the corresponding letter of a reaction at the position defined by the reaction's index (see Fig. 2.1). The

empty spaces are filled with zeroes. The string describing the whole sequence I will be $J_1 J_2 \ldots J_U$. For example, the pathway depicted in Fig. 2.2 will be described by the string

h0A000h000A0 0000N000p00p 0H0000000000

where the 'u'-reactions have been omitted.

Having two such strings S_1 and S_2 belonging to two reaction sequences C_1 and C_2, respectively, we simply define $D(C_1, C_2)$ to be the number of positions in which S_1 and S_2 differ.

A.4.2 The arrangement of coupling reactions inside a reaction chain

Let x and y denote two different types of generic reactions, C the sequence to be examined. Further, let X and Y be the number of reactions of types x and y in C, respectively. Let Y_n^+ be the number of reactions of type y occurring after the n-th position in the reaction chain C, let Y_n^- be the number of such reactions before the n-th position. Additionally, let $\{\xi_i, i = 1 \ldots X\}$ define the set containing the positions of all reactions of type x. We define as a measure of the internal ordering of the reactions x and y

$$p_{xy}(C) = \frac{1}{X \cdot Y} \sum_{i=1}^{X} \left(Y_{\xi_i}^+ - Y_{\xi_i}^- \right). \tag{A.29}$$

For example, let us calculate p_{Aa} for the last reaction sequence in Table 2.2. The only reactions playing a role for the calculation of p_{Aa} are those of type 'A' or 'a'. As there are six 'A'- and six 'a'-reactions, we have $X = Y = 6$. Extracting the relevant reactions, we get the string **AaAaaAAaAAaa**, which yields

$$p_{Aa} = \frac{1}{36} (6 + 4 + 0 + 0 - 2 - 2) = \frac{1}{6}. \tag{A.30}$$

It is easy to see that the definition (A.29) ensures that the conditions (2.36)–(2.38) are fulfilled. For example, assume, all X reactions of type x are situated before all Y reactions of type y in reaction chain C. Eq. (A.29) reduces to

$$p_{xy}(C) = \frac{1}{X \cdot Y} \sum_{i=1}^{X} Y = 1, \tag{A.31}$$

which is in accordance with Eq. (2.36).

Appendix B

Mathematical addendum to chapter 3

B.1 Maximal size of elementary networks

Theorem 2 *Let L denote the maximal number of carbon atoms in the participating compounds.*

A network which is elementary with respect to a conversion $\langle a, b\rangle$ consists of maximally $r = L - 1$ reactions.

Proof

Let $\mathbf{A}$ be a matrix with r columns. Such a matrix defines a linear map. Let $\mathrm{Im}(\mathbf{A})$ denote the range of the map defined by $\mathbf{A}$ and $\mathrm{Ker}(\mathbf{A})$ denote the kernel of the map. According to the rules of linear algebra, the following relation holds

$$\dim \mathrm{Im}(\mathbf{A}) + \dim \mathrm{Ker}(\mathbf{A}) = r. \tag{B.1}$$

Now let's consider a metabolic reaction network consisting of r reactions which is elementary with respect to the conversion $\langle a, b\rangle$. Let $\mathbf{N}$ be the stoichiometric matrix describing the system which fulfills the steady state condition (3.2). By definition, for an elementary network there exists exactly one linearly independent solution of Eq. (3.2). Consequently, the reduced stoichiometric matrix $\mathbf{N}^{(a,b)}$ which is constructed from $\mathbf{N}$ by deleting the rows corresponding to the compounds C_0, C_a and C_b has a one-dimensional kernel, i. e. $\dim \mathrm{Ker}(\mathbf{N}^{(a,b)}) = 1$. The matrix $\mathbf{N}^{(a,b)}$ has r columns and maximally $L-2$ rows with non-zero entries. Thus, $\dim \mathrm{Im}(\mathbf{N}^{(a,b)}) \leq L-2$, which yields together with relation (B.1) applied to the matrix $\mathbf{N}^{(a,b)}$:

$$r \leq L - 1 \tag{B.2}$$

which proves the theorem.

Most calculations in this work were carried out with $L = 6$ resulting in a maximal size of an elementary network of $r = 5$.

B.2 Proof of symmetry

Theorem 3 *Let L denote the maximal number of carbon atoms in the participating compounds.*

For any given network size r there exist exactly the same number of networks performing the conversion $\langle L, a\rangle$ in an elementary way as networks performing the conversion $\langle L, L-a\rangle$ in an elementary way.

In the following the theorem is proven by explicitely constructing a one to one relation mapping networks performing the conversion $\langle L, a\rangle$ into networks performing the conversion $\langle L, L-a\rangle$.

Let $\mathfrak{R}$ be the set of all reactions (of type 1 as well as type 2) as defined in section 3.1. Formally,

$$\mathfrak{R} = \{(i, j|k, l) : i \leq j \leq L, k \leq l \leq L, i \leq k\} \tag{B.3}$$

Let the bijective map $X : \mathfrak{R} \to \mathfrak{R}$ be defined as follows:

$$X(i, j|k, l) = (L-l, L-k|L-j, L-i). \tag{B.4}$$

This map is obviously bijective. Furthermore, the map is its own reverse, i. e. $X^2(i, j|k, l) = (i, j|k, l)$.

Let $\mathcal{N}$ be a network of size r which is elementary with respect to the conversion $\langle L, a\rangle$. Further, let $\mathcal{R}_1, \ldots, \mathcal{R}_r$ denote the reactions participating in the network and $\mathbf{N}$ denote the stoichiometric matrix describing the network. For matters of convenience, the stiochiometric matrix is written using its row vectors $n_0, \ldots, n_L$:

$$\mathbf{N} = \begin{pmatrix} n_0 \\ \vdots \\ n_L \end{pmatrix} \tag{B.5}$$

According to the steady state condition (3.2), there exists a vector $V = (v_0, \ldots, v_L)^T$ such that

$$\mathbf{N} \cdot V = S \tag{B.6}$$

where $S = (s_0, \ldots, s_L)^T$ with $s_a = -L$, $s_L = a$ and $s_i = 0$ for $i \neq 0, a, L$.

Let $\bar{\mathcal{N}}$ be the network consisting of the reactions $X(\mathcal{R}_1), \ldots, X(\mathcal{R}_r)$. It follows from the definition (B.4) of the map X that the stoichiometrix matrix $\bar{\mathbf{N}}$ describing this network reads

$$\bar{\mathbf{N}} = \begin{pmatrix} -n_L \\ \vdots \\ -n_0 \end{pmatrix} \tag{B.7}$$

Now, obviously the vector $\bar{V} = (-v_L, \ldots, -v_0)^T$ fulfills the condition

$$\bar{\mathbf{N}} \cdot \bar{V} = \bar{S} \tag{B.8}$$

where $\bar{S} = (\bar{s}_0, \ldots \bar{s}_L)^T$ with $\bar{s}_L = s_0$, $\bar{s}_{L-a} = -L$ and $\bar{s}_i = 0$ for $i \neq 0, L-a, L$.

Comparing this with the steady state condition (3.2) shows that the network $\bar{\mathcal{N}}$ performs the conversion $\langle L, L-a \rangle$.

Thus, a bijective map between the set of networks performing the conversion $\langle L, a \rangle$ and the set of networks performing the conversion $\langle L, L-a \rangle$ has been explicitely constructed, therefore their numbers of elements must be equal and the theorem is proven.

It should be noted that the steady state conditions (B.6) and (B.8) for the networks $\mathcal{N}$ and $\bar{\mathcal{N}}$ originate from each other by reversing the order of the rows (and inverting all signs). This can easily be used to show that the network $\bar{\mathcal{N}}$ is elementary with respect to a conversion $\langle a, b \rangle$ if and only if the network $\mathcal{N}$ is elementary with respect to this conversion (compare with the algorithm described in section 3.2.1).

B.3 Omnipotent networks

Theorem 4 *Let L denote the maximal number of carbon atoms in the participating compounds.*

Any network of size $L-1$ which is elementary to an arbitrary conversion $\langle \alpha, \beta \rangle$, $1 \leq \alpha < \beta \leq L$ can perform all conversions $\langle a, b \rangle$ with $1 \leq a < b \leq L$.

In the proof of the Theorem the compound C_0 does not play a role and therefore all stoichiometric matrices are understood without the corresponding row for this formal compound. Equivalently to appendix B.1, the notation $\mathbf{N}^{(a,b)}$ is used for the matrix which emerges from $\mathbf{N}$ by deletion of the a-th and b-th row. This notation is analogously used for vectors.

Before presenting the proof, I postpone the following Lemma which is a direct consequence from the conservation of the numbers of carbon atoms:

Lemma 1 *For any network of an arbitrary size r described by the stoichiometric matrix $\mathbf{N}$ and any vector $X \in \mathbb{R}^r$ the vector $Y = \mathbf{N} \cdot X$ fulfils the condition*

$$\sum_{i=1}^{L} iy_i = 0, \tag{B.9}$$

where the y_i denote the components of the vector Y.

Since the columns of a stoichiometric matrix represent the stoichiometries of the participating reactions, and since every single reaction of course conserves the number of carbon atoms, an analogue of relation (B.9) holds true for every single column. Thus, relation (B.9) is also true, since Y is nothing else but a linear combination of the column vectors of the stoichiometric matrix $\mathbf{N}$,.

For the proof of the Theorem, let $\mathbf{N}$ be the stoichiometric matrix of a reaction network of size $L-1$ that is elementary with respect to the conversion $\langle \alpha, \beta \rangle$.

Consequently, there exists a unique vector V such that

$$\mathbf{N} \cdot V = \mathbf{\Sigma} \equiv (\sigma_1, \ldots, \sigma_L)^T \text{ with } \sigma_\alpha = -\beta,\ \sigma_\beta = \alpha,\ \sigma_i = 0 \text{ for } i \neq \alpha, \beta \tag{B.10}$$

– see Eq. (3.2). From the proof of Theorem 2 it is known that dim $\mathrm{Ker}(\mathbf{N}^{(\alpha,\beta)}) = 1$ and Eq. (B.1) yields dim $\mathrm{Im}(\mathbf{N}^{(\alpha,\beta)}) = L-2$. This means that the range of the image of the map defined by $\mathbf{N}^{(\alpha,\beta)}$ is identical to the whole space $\mathbb{R}^{L-2}$. The consequence is the following

Lemma 2 *For all vectors $T \in \mathbb{R}^{L-2}$ there exists a vector $W \in \mathbb{R}^{L-1}$ such that $\mathbf{N}^{(\alpha,\beta)} \cdot W = T$.*

Now let's choose a conversion $\langle a, b \rangle$, $1 \leq a < b \leq L$ and let $S \equiv (s_1, \ldots, s_L)^T$ denote the corresponding conversion vector ($s_a = -b$, $s_b = a$ and $s_i = 0$ for $i \neq a, b$). Following Lemma 2, there exists a vector W, such that

$$\mathbf{N}^{(\alpha,\beta)} \cdot W = S^{(\alpha,\beta)}. \tag{B.11}$$

Hence,

$$\mathbf{N} \cdot W = S^* \equiv (s_1^*, \ldots, s_l^*)^T \text{ where } s_i^* = s_i \text{ except for } i \in \{\alpha, \beta\}. \tag{B.12}$$

Now let

$$\lambda = \frac{s_\beta - s_\beta^*}{\alpha}. \tag{B.13}$$

Then,

$$\mathbf{N}\,(W + \lambda V) = S \tag{B.14}$$

which can be seen as follows:

- for the components $i \neq \alpha, \beta$ the relation (B.14) is obviously true, because $\sigma_i = 0$ – see Eq. (B.10).
- for $i = \beta$, relation (B.14) is also true, since λ has been chosen such that $s^*_\beta + \lambda\alpha = s_\beta$ – see Eqs. (B.10) and (B.13).
- finally, Lemma 1 ensures that relation (B.14) holds true also for the last component ($i = \alpha$). Note here, that only two entries of S are non-zero.

With Eq. (B.14) it has been shown that a solution vector $(W + \lambda V)$ for the steady state condition (3.2) exists and therefore the corresponding network is able to perform the conversion $\langle a, b\rangle$. As the conversion has been arbitrarily chosen, the Theorem is proven.

B.4 The impossibility of bi-bi-Networks

Theorem 5 *There exists no network consisting entirely of bi-bi-reactions that fulfills the steady state condition (3.2).*

We will prove this theorem is two parts. In the first part, it will be shown that all bi-bi-reactions can be written as a combination of reactions out of a small "base" set consisiting of $L - 2$ reactions, with L being the maximal number of carbon atoms per molecule. In the second part, it will be shown that it is impossible to permorm any conversion $\langle a, b\rangle$ with a network consisting of exclusively these reactions.

Part one. We prove the following

Lemma 3 *All bi-bi-reactions can be replaced by a combination of reactions of the set*

$$\mathcal{B}_L = \{(1, j|2, j-1), j = 3, \ldots, L\} \tag{B.15}$$

The prove will be led by induction over the maximal number L of carbon atoms.

For $L = 3$, the statement of the lemma is true since the only existing bi-bi-reaction is $(1, 3|2, 2)$, which is included in the set $\mathcal{B}_L$.

Let us assume the statement is true for $L = l - 1$. Increasing the maximal number of carbon atoms from $l-1$ to l, the set of bi-bi-reactions is extended by reactions which are of the form $(k, l|k + i, l - i)$ where $k \leq l - 2$ and $i \leq (l + k)/2$ for $l + k$ even or $i \leq (l + k - 1)/2$ for $l + k$ odd. The reaction characterized by $k = 1$ and $i = 1$ is included in the set $\mathcal{B}_l$. All other reactions can be written as combinations makeing use only of reactions contained in $\mathcal{B}_l$. Two cases have to be considered:

1. $k = 1$: For $i > 1$, the reaction $(1, l|1 + i, l - i)$ can be written as the sequence of the two reactions $(1, l|2, l - 1)$ and $(2, l - 1|1 + i, L - i)$. The first reaction is included in $\mathcal{B}_l$. The second reaction makes use of compounds containing not more than $l - 1$ carbons. Since lemma 3 is assumed to be true for $L = l - 1$, this reaction can be written as a combination of reactions from the set $\mathcal{B}_{l-1}$.

2. $k > 1$: For $i = 1$, the reaction $(k, l|k + 1|l - 1)$ can be written as the sequence of the reaction $(1, l|2, l - 1)$ and the reverse of the reaction $(1, k + 1|2, k)$. The first reaction is included in $\mathcal{B}_l$. Since $k \leq l - 2$, the second makes use of compounds containing not more than $l - 1$ carbons. For $i > 1$, the reaction $(k, l|k + i, l - i)$ can be written as the sequence of the two reactions $(k, l|k + 1, l - 1)$ and $(k + 1, l - 1|k + i, l - i)$. By consideration of the case $i = 1$ it was shown that the first reaction can be written as a combination of reactions from $\mathcal{B}_l$. The second makes use of compounds containing not more than $l - 1$ carbons.

This completes the proof of lemma 3.

Part two. Consider a metabolic network consisting of all reactions of the set $\mathcal{B}_L$ characterized by the stoichiometric matrix $\mathbf{N}$. Since bi-bi-reactions do not involve the compound C_0, this matrix contains L rows for the compounds $C_1, \ldots C_L$. It remains to be shown that this network cannot perform any conversion $\langle a, b \rangle, 1 \leq a < b \leq L$. There are $L-2$ columns corresponding to all reactions of the set $\mathcal{B}_L$. The stoichiometric matrix $\mathbf{N}$ reads

$$\mathbf{N} = \begin{pmatrix} -1 & -1 & -1 & \cdots & -1 & -1 \\ 2 & 1 & 1 & \cdots & 1 & 1 \\ -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ 0 & 0 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \\ 0 & 0 & 0 & \cdots & 0 & -1 \end{pmatrix} \tag{B.16}$$

A network consisting entirely of bi-bi-reactions which can perform the conversion $\langle a, b \rangle$ would be characterized by the equation $\mathbf{N} \cdot V = S$ with a vector $V \neq 0$ and a vector S which is non-zero only in the a-th and b-th element – see Eq. 3.2. Accordingly, the reduced matrix $\mathbf{N}^{(a,b)}$ resulting from $\mathbf{N}$ by deletion of the a-th and b-th row must

fulfil the equation.

$$\mathbf{N}^{(a,b)} \cdot \mathbf{V} = \mathbf{0} \tag{B.17}$$

with $V \neq 0$. It can be shown that for every $1 \leq a < b \leq L$ the $(L-2, L-2)$-matrix $\mathbf{N}^{(a,b)}$ has full rank. Since in this case the only solution vector for Eq. B.17 is $V = 0$, networks consisting entirey of bi-bi-reactions cannot perform any conversion $\langle a, b \rangle$.

To prove that $\mathbf{N}^{(a,b)}$ has full rank, the row-echelon form $\tilde{\mathbf{N}}^{(a,b)}$ of this matrix (see for example (Groetsch and King 1988)) is constructed. In this process multiples of rows are added to other rows in such a way that an upper triangular matrix resuls which directly reveals its rank. As an example, this construction is demonstrated for the case $a = 1$ and $b > 3$. Note that for $1 \leq a < b \leq 3$, $\mathbf{N}^{(a,b)}$ is already an upper triangluar matrix.

Let n_{jk} denote the coefficients of the matrix $\mathbf{N}^{(1,b)}$. From eq. (B.16) it can be seen that

$$\begin{aligned}
&n_{11} = 2 \qquad n_{12} = 1, \ldots, n_{1,L-2} = 1 \\
&n_{j,j-1} = -1 \quad n_{j,j} = 1 \qquad \text{for} \quad j = 2 \ldots b-2 \\
&n_{j,j} = -1 \qquad n_{j,j+1} = 1 \qquad \text{for} \quad j = b-1 \ldots L-2 \\
&n_{jk} = 0 \text{ at all other positions} \qquad .
\end{aligned} \tag{B.18}$$

We construct the elements $\tilde{n}_{jk}$ of $\tilde{\mathbf{N}}^{(a,b)}$ as follows: $\tilde{n}_{1k} = n_{1k}$ for $k = 1, \ldots, L-2$; $\tilde{n}_{jk} = j \cdot n_{jk}$ for $j = 2, \ldots, b-2$ and $k = 1, \ldots, L-2$; $\tilde{n}_{jk} = n_{jk}$ for $j = b-1, \ldots, L-2$ and $k = 1, \ldots, L-2$.

The elements of $\tilde{\mathbf{N}}^{(a,b)}$ read

$$n_{j1} = 0, \ldots, n_{j,j-1} = 0, n_{jj} = j+1, n_{j,j+1} = 1, \ldots, n_{j,L-2} = 1. \tag{B.19}$$

The resulting matrix is an upper triangular matrix obviously having full rank. The cases $a = 2$, $a = 3$ and $a > 3$ can be shown in an analogue way.

Bibliography

Alberts, B., D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson (1994). *Molecular Biology of the Cell* (3 ed.). Garland Publishing Inc.

Albery, W. J. and J. R. Knowles (1976). Evolution of enzyme function and the development of catalytic efficiency. *Biochemistry 15*, 5631–5640.

Angulo-Brown, F., M. Santillán, and E. Calleja-Quevado (1995). Thermodynamic optimality in some biochemical reactions. *Nuovo Cimento D 17*, 87–90.

Banzhaf, W., P. Nordin, R. E. Keller, and F. D. Francone (1998). *Genetic Programming - An Introduction*. Morgan Kaufmann Publishers, San Francisco and dpunkt.verlag, Heidelberg.

Barabasi, A. L. and R. Albert (1999). Emergence of scaling in random networks. *Science 286*, 509–512.

Binder, B. and R. Heinrich (2002). Dynamic stability of signal transduction networks depending on downstream and upstream specificity of protein kinases. *BioComplexity*.

Brightman, F. A. and D. A. Fell (2000). Hypothesis: Differential feedback regulation of the mapk cascade underlies the quantitative differences in egf and ngf signalling in pc12 cells. *FEBS Letters 482*, 169–174.

Ebeling, W., A. Engel, and R. Feistel (1990). *Physik der Evolutionsprozesse*. Akademie Verlag Berlin.

Ebeling, W. and R. Feistel (1977). Stochastic theory of molecular replication processes with selection character. *Ann. Phys. 34*, 81–90.

Eigen, M. (1971). Selforganization of matter and the evolution of biological macromolecules. *Die Naturwissenschaften 58*, 465–523.

Eigen, M., J. McCaskill, and P. Schuster (1989). Molecular quasispecies. *J. Phys. Chem. 92*, 6881–6891.

Ferea, T. L., D. Botstein, P. O. Brown, and R. F. Rosenzweig (1999). Systematic changes in gene expression patterns following adaptive evolution in yeast. *Proc. Natl. Acad. Sci. USA 96*, 9721–9726.

Florkin, M. and E. H. Stotz (Eds.) (1969). *Carbohydrate Metabolism*. Elsevier, Amsterdam.

Garfinkel, D. and B. Hess (1964). Metabolic control mechanism vii. a detailed computer model of the glycolytic pathway in ascites cells. *J. Biol. Chem. 239*, 971–983.

Goldberg, D. (1989). *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley.

Groetsch, C. W. and J. T. King (1988). *Matrix methods and applications*. Prentice-Hall, Englewood Cliffs.

Heinrich, R. and E. Hoffmann (1991). Kinetic parameters of enzymatic reactions in states of maximal activity. an evolutionary approach. *J. Theor. Biol. 151*, 249–283.

Heinrich, R., H. G. Holzhütter, and S. Schuster (1987). A theoretical approach to the evolution and structural design of enzymatic networks: linear enzymatic chains, branched pathways and glycolysis of erythrocytes. *Bull. Math. Biol. 49*, 539–595.

Heinrich, R., F. Montero, E. Klipp, T. G. Waddell, and E. Meléndez-Hevia (1997). Theoretical approaches to the evolutionary optimization of glycolysis; thermodynamic and kinetic constraints. *Eur. J. Biochem. 243*, 191–201.

Heinrich, R. and S. Schuster (1996). *The Regulation of Cellular Systems*. Chapman & Hall, New York.

Heinrich, R. and I. Sonntag (1981). Analysis of the selection equations for a multivariable population model: Deterministic and stochastic solutions and discussion of the approach for populations of self-reproducing biochemical networks. *J. theor. Biol. 93*, 325–361.

Huang, F. C. and J. E. Ferrel (1996). Ultrasensitivity in the mitogen activated protein kinase cascade. *Proc. Natl. Acad. Sci. USA 93*, 10078–10083.

Jeong, H., B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási (2000). The large-scale organization of metabolic networks. *Nature 407*, 651–654.

Joshi, A. and B. Ø. Palsson (1989). Metabolic dynamics in the human red cell. i. and ii. *J. Theor. Biol. 141*, 515–545.

Joshi, A. and B. Ø. Palsson (1990). Metabolic dynamics in the human red cell. iii. and iv. *J. Theor. Biol. 142*, 41–85.

Mavrovouniotis, M. L. and G. Stepanopoulos (1990). Estimation of upper bounds for the rates of enzymatic reactions. *Chem. Eng. Commun. 93*, 211–236.

Mavrovouniotis, M. L., G. Stepanopoulos, and G. Stepanopoulos (1990). Computer-aided synthesis of biochemical pathways. *Biotechnology and Bioengineering 36*, 1119–1132.

Meléndez-Hevia, E. and A. Isidoro (1985). The game of the pentose phosphate cycle. *J. Theor. Biol. 117*, 251–263.

Meléndez-Hevia, E. and V. Torres (1988). Economy of design in metabolic pathways: further remarks on the game of the pentose phosphate cycle. *J. Theor. Biol. 132*, 97–111.

Meléndez-Hevia, E., T. G. Waddell, R. Heinrich, and F. Montero (1997). Theoretical approaches to the evolutionary optimization of glycolysis; chemical analysis. *Eur. J. Biochem. 244*, 527–543.

Michal, G. (Ed.) (1999). *Biochemical Pathways*. Spektrum Akademischer Verlag Heidelberg, Berlin.

Mittenthal, J. E., B. Clarke, T. G. Waddell, and G. Fawcett (2001). A new method for assembling metabolic networks with application to the krebs citric acid cycle. *J. theor. Biol. 208*, 361–382.

Mittenthal, J. E., A. Yuan, B. Clarke, and A. Scheeline (1998). Designing metabolism; alternative connectivities for the pentose-phosphate pathway. *Bull. Math. Biol. 60*, 815–856.

Mulquiney, P. and P. W. Kuchel (1999a). Model of 2, 3-bisphosphoglycerate metabolism in the human erythrocyte based on detailed enzyme kinetic equations: Equations and parameter refinement. *Biochem. J. 342*, 581–596.

Mulquiney, P. and P. W. Kuchel (1999b). Model of 2, 3-bisphosphoglycerate metabolism in the human erythrocyte based on detailed enzyme kinetic equations: Computer simulation and metabolic control analysis. *Biochem. J. 342*, 597–604.

Nuño, J. C., I. Sanchez-Valdenebro, C. Pereziratxeta, E. Meléndez-Hevia, and F. Montero (1997). Network organization of cell metabolism; monosaccharide interconversion. *Biochem. J. 324*, 103–111.

Pettersson, G. (1992). Evolutionary optimization of the catalytic efficiency of enzymes. *Eur. J. Biochem. 206*, 289–295.

Rapoport, T., R. Heinrich, and S. M. Rapoport (1976). The regulatory principles of glycolysis in erythrocytes in vivo and in vitro. a minimal comprehensive model describing steady states, quasi-steady states and time dependent processes. *Biochem. J. 154*, 449–469.

Rechenberg, I. (1989). *Evolution strategy: Nature's way of optimization*, Volume 47 of *Lecture notes in Engineering*. Springer, Berlin.

Rizzi, M., M. Baltes, U. Theobald, and M. Reuss (1997). In vivo analysis of metabolic dynamics in saccharomyces cerevisiae: Ii. mathematical model. *Biotechnol. Bioeng. 55*, 592–608.

Sachs, L. (1992). *Angewandte Statistik*. Springer-Verlag, Berlin, Heidelberg.

Saier Jr., M. H. (Ed.) (2002). *The Bacterial Phosphotransferase System*. Horizon Scientific Press.

Schilling, C. H. and B. Ø. Palsson (1998). The underlying pathway structure of biochemical networks. *Proc. Natl. Acad. Sc. 95*, 4193–4198.

Schilling, C. H., S. Schuster, B. Ø. Palsson, and R. Heinrich (1999). Metabolic pathway analysis: Basic concepts and scientific applications in the post-genomic era. *Biotechn. Prog. 199*, 45–61.

Schuster, S., T. Dandekar, and D. A. Fell (1999). Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends in Biotechnology 17*, 53–60.

Schuster, S. and C. Hilgetag (1994). On elementary flux modes in biochemical reaction systems at steady state. *J. Biol. Syst. 2*, 165–182.

Stephani, A. and R. Heinrich (1998). Kinetic and thermodynamic principles determining the structural design of atp-producing systems. *Bull. Math. Biol. 60*, 505–543.

Stephani, A., J. C. Nuño, and R. Heinrich (1999). Optimal stoichiometric design of atp-producing systems as determined by an evolutionary algorithm. *J. Theor. Biol. 199*, 45–61.

Strogatz, S. H. (2001). Exploring complex networks. *Nature 410*, 268–276.

Stryer, L. (1988). *Biochemistry*. Freeman and Company, New York.

Teusink, B., J. Passarge, C. A. Reijenga, E. Esgalhado, C. C. van der Weijden, M. Schepper, M. C. Walsh, B. M. Bakker, K. van Dam, H. V. Westerhoff, and J. L. Snoep (2000). Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? testing biochemistry. *Eur. J. Biochem. 267*, 5313–5329.

Teusink, B., M. C. Walsh, K. Van Dam, and H. V. Westerhoff (1998). The danger of metabolic pathways with turbo design. *Trends in Biochemical Sciences 23*, 162–169.

Varma, A. and B. Ø. Palsson (1993). Metabolic capabilities of *escherichia coli*: I. synthesis of biosynthetic precursors and cofactors. *J. theor. Biol. 165*, 477–502.

Venables, W. N. and B. D. Ripley (1998). *Modern Applied Statistics with S-PLUS* (2 ed.)., Chapter 13. Springer, New York.

Waddell, T. G., P. Repovic, E. Meléndez-Hevia, R. Heinrich, and F. Montero (1999). Optimization of glycolysis: new discussions. *Biochem. Educ. 27*, 12–13.

Wagner, A. and D. A. Fell (2001). The small world inside large metabolic networks. *Proc. Royal Soc. 268*, 1803–1810.

Werner, A. and R. Heinrich (1985). A kinetic model for the interaction of energy metabolism and osmotic states of human erythrocytes. analysis of the stationary "in vivo" state and of time dependent variations under blood preservation conditions. *Biomed. Biochim. Acta. 44*, 185–212.

Wilhelm, T., E., E. Hoffmann-Klipp, and R. Heinrich (1994). An evolutionary approach to enzyme kinetics: optimization of ordered mechanisms. *Bull. Math. Biol. 56*, 65–106.

Wolf, J. and R. Heinrich (2000). Effect of cellular interaction on glycolytic oscillations in yeast: A theoretical investigation. *Biochem. J. 345*, 321–334.

Wolf, J., H. Y. Sohn, R. Heinrich, and H. Kuriyama (2001). Mathematical analysis of a mechanism for autonomous metabolic oscillations in continuous culture of saccharomyces cerevisiae. *FEBS Lett. 499*(3), 230–234.

Lebenslauf

Persönliche Daten	Oliver Ebenhöh Seelower Str. 9 10439 Berlin Tel. 030-44731621 E-mail: oliver.ebenhoeh@rz.hu-berlin.de
Schulausbildung	Grundschule, Orientierungsstufe und Gymnasium in Oldenburg (Oldb.) Abitur Mai 1989 am Gymnasium Cäcilienschule Oldenburg
Universitätsausbildung	Physikstudium an der Ruprecht-Karls-Universität Heidelberg, WS 1989 / 90 – WS 1995 / 96 Lehramtsstudium Mathematik / Physik mit Erweiterungsfach Pädagogik, zusätzlich SS 1996
Hochschulabschlüsse	Diplom Physik, Dezember 1995 1. Staatsexamen Mathematik / Physik, April 1996 Pädagogikum (Zusatzqualifikation zum 1. Staatsexamen), Juli 1996
Berufserfahrung	Tutorenstelle (Übungsgruppenleitung) am Institut für Mathematik an der Universität Heidelberg, WS 1991 / 1992 – SS 1992 und WS 1993 / 1994 – WS 1995 / 96 IT Consultant bei Control Data GmbH., Frankfurt (Main), September 1996 bis Juni 1998 Wissenschaftlicher Mitarbeiter an der Humboldt-Universität zu Berlin, Institut für Biologie, Arbeitsgruppe theoretische Biophysik, seit August 1998
Beginn der Promotionsarbeit	15. August 1998
Auslandsaufenthalte	Auslandsstudium an der University of Aberdeen, Scotland, September 1992 bis Juni 1993 Traineeprogramm der Firma Control Data in Minneapolis, USA, September bis Dezember 1996 Forschungsaufenthalt an der Vrije Universiteit Amsterdam, Niederlande, Juni / Juli 2001
Fremdsprachen	Englisch, verhandlungssicher Spanisch, fortgeschrittene Kenntnisse Russisch, Konversation Niederländisch, Konversation

Berlin, den 9. Januar 2003

Publikationsliste

Ebenhöh, O. and R. Heinrich (1999). Structural analysis of ATP and NADH producing systems using optimisation principles. In: *Theory and Mathematics in Biology and Medicine, 4th EMSTB meeting.* Amsterdam, Netherlands.

Ebenhöh, O. and R. Heinrich (2000). Reconstruction of the stoichiometry of ATP and NADH-producing systems using evolutionary algorithms. In: *BioThermoKinetics - Animating the Cellular Map.* Eds. J.-H. S. Hofmeyr, J. M. Rohwer and Snoep, J. L., Stellenbosch, South Africa.

Ebenhöh, O. and R. Heinrich (2001). Evolutionary Optimization of Metabolic Pathways. Theoretical Reconstruction of the Stoichiometry of ATP and NADH producing systems. *Bull. Math. Biol. 63*, 21-55.

Ebenhöh, O. and R. Heinrich (2002). Structural design of metabolism. In: *International Workshop Cell Systems Biology.* Berlin, Germany.

Ebenhöh, O. and R. Heinrich (2002). Stoichiometric Design of Metabolic Networks: Multifunctionality, Clusters, Optimisation, Weak and Strong Robustness. Submitted to *Bull. Math. Biol.*

Danksagung

An erster Stelle möchte ich Herrn Prof. R. Heinrich für seine hervorragende Betreuung und das Interesse, das er meiner Arbeit entgegengebracht hat, danken. Er hat es immer geschafft, mir durch seine kritischen Fragen neue Denkanstöße zu geben und stand mir auch in schwierigen Phasen der Arbeit stets mit Rat und Tat zur Seite.

Herrn Dr. Stefan Schuster danke ich für seine zahlreichen und hilfreichen Hinweise und Verbesserungsvorschläge. Ihm und Herrn Prof. W. Ebeling danke ich für ihr Interesse an meiner Arbeit und für die Bereitschaft, diese zu begutachten.

Ein wesentlicher Faktor für das Gelingen dieser Arbeit war natürlich die angenehme Arbeitsathmosphäre, die ich in dieser Arbeitsgruppe genießen durfte. Dafür möchte allen Mitgliedern - insbesondere auch den ehemaligen - meinen Dank aussprechen.

Besonders hervorheben möchte ich Antonio Politi und Bernd Binder für ihre aufopferungsvolle Bereitschaft, meine Arbeit korrekturzulesen.

Von den ehemaligen Mitgliedern möchte ich besonders Dr. Amadeus Stephani hervorheben, der es geschafft hat, in der frühen Phase der Doktorarbeit in mir die Begeisterung für evolutionäre Algorithmen zu wecken; ebenso Dr. Stephan Frickenhaus, von welchem ich durch die unterhaltsamen und interessanten Diskussionen sehr profitiert habe, insbesondere hinsichtlich computertechnischer Fragen und Programmiertechniken. Für die zahlreichen anregenden Unterhaltungen sowie dafür, dass sie stets ein offenes Ohr hatte, möchte ich mich ganz besonders bei Dr. Edda Klipp bedanken.

Ein ganz besonderer Dank gilt Herrn Dr. Robert Arndt, dessen technisches Geschick für unsere Arbeitsgruppe ein wahrer Segen ist.

Meiner Frau Angelika, die an meiner Seite Höhen und Tiefen während dieser Arbeit miterlebt hat, bin ich zutiefst für ihre nahezu unermessliche Geduld mit mir dankbar. Sie hat mich stets in meinen Vorhaben unterstützt.

Abschließend möchte ich meinen Eltern meine tiefe Dankbarkeit aussprechen, ohne deren Unterstützung so manche Lebensabschnitte ungleich schwerer gewesen wären.

Erklärung

Ich versichere hiermit, die vorliegende Arbeit selbständig und ausschließlich unter Verwendung der angegebenen Mittel und ohne unerlaubte Hilfen angefertigt zu haben.

Berlin, den 9. Dezember 2002

Oliver Ebenhöh